Vera L. Beck
Editor

Linear Regression

Models, Analysis and Applications

Mathematics Research Developments

NOVA

MATHEMATICS RESEARCH DEVELOPMENTS

LINEAR REGRESSION

MODELS, ANALYSIS AND APPLICATIONS

No part of this digital document may be reproduced, stored in a retrieval system or transmitted in any form or by any means. The publisher has taken reasonable care in the preparation of this digital document, but makes no expressed or implied warranty of any kind and assumes no responsibility for any errors or omissions. No liability is assumed for incidental or consequential damages in connection with or arising out of information contained herein. This digital document is sold with the clear understanding that the publisher is not engaged in rendering legal, medical or any other professional services.

MATHEMATICS RESEARCH DEVELOPMENTS

Additional books in this series can be found on Nova's website under the Series tab.

Additional e-books in this series can be found on Nova's website under the eBooks tab.

ANALYTICAL CHEMISTRY AND MICROCHEMISTRY

Additional books in this series can be found on Nova's website under the Series tab.

Additional e-books in this series can be found on Nova's website under the eBooks tab.

MATHEMATICS RESEARCH DEVELOPMENTS

LINEAR REGRESSION

MODELS, ANALYSIS AND APPLICATIONS

VERA L. BECK
EDITOR

nova
science publishers
New York

Copyright © 2017 by Nova Science Publishers, Inc.

All rights reserved. No part of this book may be reproduced, stored in a retrieval system or transmitted in any form or by any means: electronic, electrostatic, magnetic, tape, mechanical photocopying, recording or otherwise without the written permission of the Publisher.

We have partnered with Copyright Clearance Center to make it easy for you to obtain permissions to reuse content from this publication. Simply navigate to this publication's page on Nova's website and locate the "Get Permission" button below the title description. This button is linked directly to the title's permission page on copyright.com. Alternatively, you can visit copyright.com and search by title, ISBN, or ISSN.

For further questions about using the service on copyright.com, please contact:
Copyright Clearance Center
Phone: +1-(978) 750-8400 Fax: +1-(978) 750-4470 E-mail: info@copyright.com.

NOTICE TO THE READER

The Publisher has taken reasonable care in the preparation of this book, but makes no expressed or implied warranty of any kind and assumes no responsibility for any errors or omissions. No liability is assumed for incidental or consequential damages in connection with or arising out of information contained in this book. The Publisher shall not be liable for any special, consequential, or exemplary damages resulting, in whole or in part, from the readers' use of, or reliance upon, this material. Any parts of this book based on government reports are so indicated and copyright is claimed for those parts to the extent applicable to compilations of such works.

Independent verification should be sought for any data, advice or recommendations contained in this book. In addition, no responsibility is assumed by the publisher for any injury and/or damage to persons or property arising from any methods, products, instructions, ideas or otherwise contained in this publication.

This publication is designed to provide accurate and authoritative information with regard to the subject matter covered herein. It is sold with the clear understanding that the Publisher is not engaged in rendering legal or any other professional services. If legal or any other expert assistance is required, the services of a competent person should be sought. FROM A DECLARATION OF PARTICIPANTS JOINTLY ADOPTED BY A COMMITTEE OF THE AMERICAN BAR ASSOCIATION AND A COMMITTEE OF PUBLISHERS.

Additional color graphics may be available in the e-book version of this book.

Library of Congress Cataloging-in-Publication Data

ISBN: 978-1-53612-015-8 (eBook)

Published by Nova Science Publishers, Inc. † New York

CONTENTS

Preface		**vii**
Chapter 1	Weighting and Transforming Data in Linear Regression *Julia Martín, Alberto Romero Gracia and Agustín G. Asuero*	**1**
Chapter 2	Regression through the Origin *Julia Martín and Agustín G. Asuero*	**69**
Chapter 3	Linear Regression for Interval-Valued Data in *Kc* (R) *Yan Sun and Chunyang Li*	**117**
Chapter 4	Linear Regression versus Non-Linear Regression in Mathematical Modeling of Adsorption Processes *Gabriela-Nicoleta Moroi*	**149**
Index		**179**

PREFACE

Chapter One addresses the importance of weighted linear regression in fitting straight lines. In Chapter Two, the authors cover the homocedastic condition, i.e., variance of y's independent of x, errors of y's accumulative, the heterocedastic case, i.e., variance or standard deviation proportional to x values, respectively, and orthogonal regression (error in both axes). The chapter also covers topics such as prediction (using the regression line in reverse), leverage, goodness of fit, comparison between models with and without intercept, uncertainty, polynomial regression models without intercept, and an overview of robust regression through the origin. Chapter Three focuses on linear regression for interval-valued data within the framework of random sets, and proposes a new model that generalizes a series of existing ones. Chapter Four provides an investigation on modeling of adsorption of heavy metal ions onto surface-functionalized polymer beads. Linear and non-linear regressions were employed for each of the isotherm models considered to describe the equilibrium data. To reliably assess model validity, various error functions (whose mathematical expressions contain the number of experimental measurements, the numbers of independent variables and parameters in the regression equation as well as the measured and predicted equilibrium adsorption capacities) were used.

Chapter 1 - Improper parameter estimation is achieved when non constant variance (heterocedasticity) is ignored. For this reason the importance of weighted linear regression in fitting straight lines is stressed in this chapter. A

number of issues are thus addressed concerning random error, noise and variance modelling when precision varies as the values of x (e.g., concentration) increase. The use of data transformation and weighted least squares regression are two main solutions to deal with the heterocedasticity problem. Non-linear terms may be introduced into the frame of linear regression by transforming variables. Fitting is improved in this way and necessary assumptions involved in least squares method such as homocedastivity (constant variance) are thus satisfied. The following topics concerning transformations are covered on this context: reasons to carry out, simplification of relationships, model linearization, variance stabilization and weighting transformation data. Box-Cox transformation topic has also received a distinctive attention. Applications (weighting, transformation and Box-Cox method) from a variety of fields (analytical, biochemical, clinical, environmental and pharmaceutical) are summarized in tabular form. The chapter is based on two previous reviews published by the authors in Critical Reviews in Analytical Chemistry (2007, 37(3) 143-172 and 2011, 41(1), 36-69).

Chapter 2 - Regression through the origin, a very interesting topic, has usually received a scarce attention in the bibliography. This model is also known as the no-intercept model. It is applied because of subject matter theory or either when other physical and material considerations are necessary to taken into account. An intensive bibliographical search has been carried out with the purpose of gathering the literature on the subject, which is widely scattered. Some about one hundredth and thirty references have been compiled, comprising about twenty monographs and fifty scientific journals, from varying fields, e.g., analytical, biological, clinical, chemometrical, educational, environmental, pharmaceutical, physico-chemical, and statistical. The authors will dealt systematically with the homocedastic condition, i.e., variance of y's independent of x, errors of y's accumulative, the heterocedastic case, i.e., variance or standard deviation proportional to x values, respectively, and orthogonal regression (error in both axes). The chapter also covers topics such as prediction (using the regression line in reverse), leverage, goodness of fit, comparison between models with and without intercept, uncertainty, polynomial regression models without intercept, and an overview of robust regression through the origin.

Chapter 3 - In the recent scientific research, data are increasingly taking on new formats such as sets, lists, and histograms. Among these, a particular type that is frequently encountered is interval-valued data, which refers to collection of observations in the form of intervals. Examples are daily [min, max] temperature, spatially [low, high] elevation, range of a group of individual observations, among many others. Linear regression as a fundamental tool of statistical analysis has been increasingly investigated for extensions to accommodate interval-valued data. Various models and methods have been proposed and studied in the last decades. However, issues such as interpretability and computational feasibility still remain. Especially, a commonly accepted mathematical foundation is largely underdeveloped, compared to the demand of applications. In this chapter, the authors focus on linear regression for interval-valued data within the framework of random sets, and propose a new model that generalizes a series of existing ones. By proposing the authors' model, the authors continue to build up the theoretical framework that deeply understands the existing models and facilitates future developments. In particular, the authors establish important properties of the model in the space of compact convex subsets of R, analogous to those for the classical linear regression. Additionally, the authors carry out theoretical investigations into the least squares estimation that is widely used in the literature. It is shown that the least squares estimator is asymptotically unbiased. A simulation study is presented that supports the authors' theorems, and an application to a climate data set is demonstrated.

Chapter 4 - In mathematical modeling of adsorption processes, linear and/or non-linear regression analysis may be employed. In adsorption isotherm modeling, non-linear regression has lately been reported by some authors to provide a better fit to experimental data than linear regression. Isotherm models used in describing the adsorption systems, criteria selected to evaluate isotherm model validity as well as modeling results are comparatively discussed. In the authors' investigation on modeling of adsorption of heavy metal ions onto surface-functionalized polymer beads, linear and non-linear regressions were employed for each of the isotherm models considered to describe the equilibrium data. To reliably assess model validity, various error functions (whose mathematical expressions contain the number of experimental measurements, the numbers of independent variables and parameters in the regression equation as well as the measured and predicted

equilibrium adsorption capacities) were used. The modeling results obtained by employing the two regression methods were compared. For the adsorption of each metal ion species, it was revealed that (a) for a particular isotherm model, the regression providing the best fit is linear, non-linear or both linear and non-linear, and (b) the order of isotherm model validities indicated via linear regression is the same with that shown by non-linear regression.

In: Linear Regression
Editor: Vera L. Beck
ISBN: 978-1-53611-992-3
© 2017 Nova Science Publishers, Inc.

Chapter 1

WEIGHTING AND TRANSFORMING DATA IN LINEAR REGRESSION

Julia Martín, Alberto Romero Gracia and Agustín G. Asuero*
Department of Analytical Chemistry, Faculty of Pharmacy,
The University of Seville, Seville, Spain

ABSTRACT

Improper parameter estimation is achieved when non constant variance (heterocedasticity) is ignored. For this reason the importance of weighted linear regression in fitting straight lines is stressed in this chapter. A number of issues are thus addressed concerning random error, noise and variance modelling when precision varies as the values of x (e.g., concentration) increase. The use of data transformation and weighted least squares regression are two main solutions to deal with the heterocedasticity problem. Non-linear terms may be introduced into the frame of linear regression by transforming variables. Fitting is improved in this way and necessary assumptions involved in least squares method

* Corresponding Author address: Agustín G. Asuero, Department of Analytical Chemistry, Faculty of Pharmacy, University of Seville, Seville, Spain.

such as homocedastivity (constant variance) are thus satisfied. The following topics concerning transformations are covered on this context: reasons to carry out, simplification of relationships, model linearization, variance stabilization and weighting transformation data. Box-Cox transformation topic has also received a distinctive attention. Applications (weighting, transformation and Box-Cox method) from a variety of fields (analytical, biochemical, clinical, environmental and pharmaceutical) are summarized in tabular form. The chapter is based on two previous reviews published by the authors in Critical Reviews in Analytical Chemistry (2007, 37(3) 143-172 and 2011, 41(1), 36-69).

Keywords: least squares method, weighting, transforming data

INTRODUCTION

Simple linear regression assumes the homocedasticity property (regular or uniform variance) and is widely applied in natural and physical sciences (Asnin, 2016; Olivieri, 2015, Lavagnini and Magno, 2007; Sayago et al., 2004; de Levie, 2000; Asuero and Gonzalez, 1989; Meites, 1979). Plots of residuals against fitted values or versus x values allow checking equal variance assumption, though it is much better to have replications (Sayago et al., 2004). Residuals are the differences, in the y-direction, between the experimental points and the corresponding fitted values, giving a minimum sum of their squares. A complete analysis in regression diagnostic requires a thorough examination of residuals. Residuals corresponding to correct fitted models should confirm the assumptions inherent in a regression analysis or failing to deny them (Meloun and Militkí, 2011; Bates and Watts, 2007; Asuero et al. 2006; Belloto and Sokolovski, 1995; Phillips et al., 1990; Ellis and Duggleby, 1978). Residuals should be randomly distributed (with equal number of plus an minus sign) when the variables are related (Miller and Miller, 2010) by means of a linear relationship (with the error symmetrically distributed).

A plot of residuals allows checking for systematic deviation between data and model, for example: i) a curvilinear pattern (higher order term to

accommodate curvature is needed), ii) systematic descending or ascending linear trend (additional terms needed), iii) fun-shaped residual pattern (inappropriate constant variance assumption), or iv) time order analysis (time effect). Regression models are used in assay development (Aarons et al., 1987), in enzymatic kinetics and pharmacokinetics, calibration, recovery studies and comparison methods, and many other pharmaceutical, biological, and chemical applications (Asuero and Bueno, 2011; Davidian, 1990). The assumption of variance homogeneity (homocedasticity) to describe a relationship between a dependent (response) variable Y and an independent (predictor) variable x usually does not hold (Tellinghuisen 2009b; Tellinghuisen, 2007; Asuero and Gonzalez, 2007); being patent instead irregular or heterogeneous variance (heterocedastic condition). Mass, substrate concentration, or temperature (Davidian, 1990), may be the predictors. Reaction rate, radioactive count, peak area or another physical property, are examples of single responses. Perform a weighted least squares regression analysis or transforming the data (Asuero and Bueno, 2011; Asuero and Gonzalez, 2007) are the two main solutions to the heterocedasticity problem.

In chemical analysis non linear calibration curves are sometimes apparent, in techniques such as liquid chromatography-mass spectrometry (matrix related non linearity effects) or atomic absorption spectrophotometry (Asnin, 2016; Mermet, 2010). In fact, in most of real problems the response function moves away from linearity as concentration values increase in a large way (end of the calibration curve). Carrying out a transformation in one or in the two variables is a mean of simplifying a non linear relationship. Keeping the model as simple as possible (minimum number of parameters fitting data at hand) is the better choice in agreement with Occam's razor. "*Non sunt multiplicanda entia praeter necessitaten*" (Bates and Watts, 2007; Garfinkel and Fegley, 1984).

Transformations to stabilize variance and to achieve normality often go hand by hand and it often happens that both assumptions are almost satisfied after carrying out an appropriate transformation. In any case as stated by Acton (1959) "*the gods who favour statisticians have frequently ordained that the world be well behaved, and so we often find that a*

transformation to obtain one of these desiderata in fact achieves them all (well, almost achieve them)." Note that dealing with fitting no model is perfect; some models suit better than others. The aim of this contribution is to underline the significance of weighting and transforming data subject in fitting straight lines models to data. Some selected examples from a variety of fields are taken from the literature and shown in tabular form in order to illustrate this book chapter.

WEIGHTING DATA

The underlying idea in weighted least squares is to give utmost importance to the most accurate data. The weighted least squares procedure entails in (Sayago et al., 2004) minimizing the weighted residuals. The benefit obtained by applying weighted least squares procedure is greater the greater the distance from homoscedasticity (Zorn et al., 1977). Table 1 compiles some formulas, which allow calculating statistics for weighted linear regression (WLR) in those cases in which data are replicated. The number of replicates required by weighted least squares is greater than the ones required by ordinary least squares. A number of factors such as the cost of calibration, standards and reagents, or the time required to perform the measurements, make, in practice, difficult to obtain the level of replicates required. However, unequal weights can also be estimated without being performed, as we will see later.

In weighted least-squares procedure each observation is characterized by a weighting value w_i (measure of the information which contains) proportional (Deming, 1943) to the inverse of the y_i.

By definition (Deming, 1943)

$$w_f = \frac{\sigma_0^2}{\sigma_f^2} \tag{1}$$

σ_0^2 is the variance of a function of unit weight (Connors, 1987), a proportionality factor. Let *f* be the mean of n_i observations $y_{i1}, y_{i2}, \ldots y_{ini}$ (random variates) from a population of standard deviation σ_0. Then we get

$$\sigma_{\bar{y}_i}^2 = \frac{\sigma_0^2}{n_i} \tag{2}$$

for the variance of $\bar{y}_i$, being its weighting factor according to Eqn. (1)

$$w_{\bar{y}_i} = \frac{\sigma_0^2}{\sigma_{\bar{y}_i}^2} = \frac{\sigma_0^2}{\frac{\sigma_0^2}{n_i}} = n_i \tag{3}$$

Table 1. Formulae for calculating statistics for weighted linear regression with replication data (Asuero and González, 2007)

Equation:	**Slope:**
$\hat{\bar{y}}_i = a_0 + a_1 x_i$	$a_1 = S_{XY} / S_{xx}$
Mean responses	Intercept
$\bar{y}_i = \left(\sum y_{iv}\right) / n_i$	$a_0 = \bar{\bar{y}} - a_1 \bar{x}$
Residual sum of squares	Weighted residuals
$SSE = \sum w_{\bar{y}_i} \left(\bar{y}_i - \hat{\bar{y}}_i\right)^2$	$w_{\bar{y}_i}^{1/2} \left(\bar{y}_i - \hat{\bar{y}}_i\right)$
Mean	Correlation coefficient
$\bar{x} = \sum w_{\bar{y}_i} x_i / \sum w_i$	$r = S_{XY} / \sqrt{S_{XX} S_{YY}}$
$\bar{\bar{y}} = \sum w_{\bar{y}_i} \bar{y}_i / \sum w_i$	Standard errors
Sum of squares about the mean	$s_{\bar{y}/x}^2 = \frac{SSE}{n-2} = \frac{S_{YY} - a_1^2 S_{XX}}{n-2}$
$S_{XX} = \sum w_{\bar{y}_i} \left(x_i - \bar{x}\right)^2$	$s_{a_0}^2 = s_{\bar{y}/x}^2 \left(\sum w_{\bar{y}_i} x_i^2\right) / \left(S_{XX} \sum w_{\bar{y}_i}\right)$
$S_{YY} = \sum w_{\bar{y}_i} \left(y_i - \bar{\bar{y}}\right)^2$	$s_{a_1}^2 = s_{\bar{y}/x}^2 / S_{XX}$
$S_{XY} = \sum w_{\bar{y}_i} \left(x_i - \bar{x}\right)\left(y_i - \bar{\bar{y}}\right)$	$\mathrm{cov}(a_0, a_1) = -\bar{x}\, s_{\bar{y}/x}^2 / S_{XX}$

We get for the weights of the k means of $w_1=n_1$, $w_2=n_2,\ldots w_k=n_k$ (as single observations have unit weight). Weights are then depending on the arbitrary factor σ_0^2 (they are relative, no absolute values).

The variance of a single observation would be σ_0^2 / w_i if the weight of each n_i original variable was w_i. If the n_i original variables were each of weight w_i (instead of unity), in this case

$$w_{\bar{y}_i} = \frac{\sigma_0^2}{\frac{\sigma_0^2}{n_i w_i}} = n_i w_i \tag{4}$$

or

$$\sigma_{\bar{y}_i}^2 = \frac{\sigma_0^2}{n_i w_i} \tag{5}$$

The weight of the mean $\overline{y}$ is as before just n times the weight of a single observation.

Let $\overline{y}_i$ now the mean value of n_i observations taken from a population of standard deviation σ_i, with single observation varying precision. Then, the variance of y_i will be σ_i^2 / n_i. Then we get in this case

$$w_{\bar{y}_i} = \frac{\sigma_0^2}{\frac{\sigma_i^2}{n_i}} = n_i \frac{\sigma_0^2}{\sigma_i^2} \tag{6}$$

Thus, weighting factors must consider both the number of replicates and the variance of each given point.

The influence of the weighting procedure in parameter estimation is depends on the nature of the experimental data set. Note that if

experimental points are not properly weighted or they are incorrectly calculated wrong results may be obtained in weighting.

Dealing with kinetic enzymatic data, the role of intuition may be of vital importance (Reigh et al., 1972).

Several kinds of weighting factors may be envisaged (Asuero and Bueno, 2011; Asuero and Gonzalez, 2007; Chow and Liu, 1995, Asuero and González, 1989; Connors, 1987, Jurs, 1986; Meites, 1979), according to the characteristics of a given data set:

(a) *Absolute Weights*. Equal weighting factors are assumed for all the points; i.e., $w_i = 1$.

(b) *Statistical Weights*. Replication for each calibration data point is required to estimate the reciprocal of variance, which prevents its application in routine practice (Mullins, 2003). For this reason, empirical weights based on x-variable (i.e., concentration) or y-variable (i.e., response) may be used as approximations, i.e., weights such as $1/x^{0.5}$, $1/x$, $1/x^2$, $1/y^{0.5}$, $1/y^2$ (Almeida et al., 2003). In those cases in which the variance of residuals decrease with x, we may also apply:

$$w_i = \frac{1}{x_{\max} - x_i} \tag{7}$$

(c) *Instrumental Weights*. Making a small number of replicates allows assigning individual weights to data points (Asuero and Gonzalez, 1989; de Levie, 2001; de Levie, 1986). However, we may assume a functional relationship between variance and the predictor (independent) variable, when there is no enough replicates (Baumann, 1997). In fact, heterocedasticity usually implies variance to be related to the expected value of the response by means of a functional relationship (Bayne and Rubin, 1986).

(d) *Transformation-Dependent Weights*. A transformation may be sometimes carried out (in one or both variables) to obtain a straight line function from a (intrinsically linear) non linear relationship (Rawlings et al., 1998; Tomassone et al., 1983).

Some assumptions theoretically necessary in applying a regression analysis may be not satisfied by the transformed data as the relative magnitudes of the errors are affected thorough the plot. Thus non-linear homocedastic data may be transformed into a linear straight line relationship with heterocedastic errors. If z_i experimental data values turn into transformed linearized data y_i (de Levie, 2012; Asuero and Bueno, 2011; Asuero and González, 1989; de Levie, 1986) the weighting factor w_i ($\sigma_0^2 = 1$) is given by

$$w_i = \left(\frac{1}{\frac{\partial y}{\partial z}} \right)^2 \tag{8}$$

Weighted least squares method is not always problem-free. Even powers of measured (untransformed signal) values, i.e., z^2 or z^4 resulting from Eqn. (8) are always positive even in those cases where the corresponding mean values average zero (de Levie, 2000). This implies that random errors in small signal regions can contribute substantially to the sum of squares distorting the analysis. The weights must also be transformed to keep the appropriate relationship (Jurs, 1970) between the weights and the points being fitted. The random error propagation law (Tellinghuisen, 2015; Tellinghuisen, 2001; Asuero and González, 1988) when applies to a function y=f(z) gives (taking σ_0^2 =1)

$$\sigma_y^2 = \sigma_z^2 \left(\frac{\partial y}{\partial z} \right)^2 \tag{9}$$

and then we get

$$w_y = w_z \left(\frac{\partial y}{\partial z} \right)^{-2} \tag{10}$$

That is, in addition to the weighting corresponding to the zi individual data point measurements, $w_z = 1/\sigma_z^2$, the transformation-dependent weighting, w_y has to be used. An overview of distinct kind of weights may be seen in Table 2.

Table 2. Some kinds of weights* (Asuero and González, 2007)

Kind	Weight	Authors
Absolute weights	1	Jurs, 1986
Statistical weights	$\frac{1}{y_i}$	Johnson, 1980; Jurs, 1986
Assumption of constant percentage error	$\frac{1}{y_i^2}$	Anderson and Snow, 1967; Smith and Mathews, 1967
Instrumental weights	$\frac{1}{s_i^2}$	Jurs, 1986
Transformation-dependent weights	$\frac{1}{(\frac{\partial y}{\partial z})^2}$	de Levie, 1986 Meites, 1979
Mixed instrumental transformation depending weights	$\frac{1}{s_z(\frac{\partial y}{\partial z})^2}$	de Levie, 1986 Meites, 1979
* s_i^2 is the estimate of σ_i^2		

RANDOM ERRORS AND NOISE

Noise is a source of random errors, which play a vital role in analytical chemistry as well as in the analysis of experimental results (Rudnyi, 1996; Prudnikov and Shapkina, 1984). Noise may be dependent of i) signals; ii) concentration; iii) other factors (Sun et al., 1994; Garden et al., 1980;

Rothman et al., 1975; Ingle, 1974; Pardue et al., 1974; Ingle and Crouch, 1972; Winefordner et al., 1970). The precision of intensity measurements in spectrochemical analysis (Klockenkämper and Bubert, 1986; Bubert and Klockenkämper, 1983) can be affected by three kinds of noise, namely, slot noise, flicker noise and detector noise. The rate and amount of ions reaching the detector are the origins of the shot noise, which follow a Poisson statistics. The process of nebulization as well as fluctuations related with the source is the origin of flicker noise, which is proportional to the signal magnitude. Detector and electronics are involved in the dark count noise. So the error total is given by

$$s_t = \sqrt{s_{shot}^2 + s_{flic}^2 + s_{\det}^2} \tag{11}$$

The sources of the three kinds of errors are summarized in Table 3.

Table 3. Likely analytical causes of the three types of signal errors (Steliopoulos et al., 2006; Lavagnini et al., 2004; Kirkup and Mulholland, 2004; van Loco et al., 2003; de Galán et al., 1985; Kemp, 1985)

Constant Error	Proportional error	Quadratic error
Sample turbidity Reagent absorbance Nonspecificity Zero error/drift Carryover Contamination	Volumetric error Gravimetric error Incomplete separation or derivation Matrix evaporation Error in time processed	Decay/dissociation of product Reagent depletion Instrumental non-linearity • matrix-related non-linearity in CG-MS • purge and trap GC-MS • electron capture detector

MODELLING THE VARIANCES AS A FUNCTION OF THE DEPENDENT OR INDEPENDENT VARIABLE

It is possible to model the variance in function of the response given its smoothing variability behaviour through the response level range (modelled either as a function of *x*, or a function of *y*) (Tellinghuisen 2010b; Tellinghuisen, 2009d; Baumann, 1997; Davidian, 1990). As a matter of fact, variance is related to the mean (or to other parameters or variables) when dealing with many physico-chemical properties (Table 4). A varying number of functions have been devised to estimate variances (Tellinghuisen, 2005a; Sadray et al., 2003; Hwang, 1994; O'Connell et al., 1993; Davidian, 1990) (Table 4).

A simple approach follows (Rodbard, et al., 1976; Rodbard and Frazier, 1975)

$$\log\left[Var(Y_{ij})\right]^{1/2} = \log\sigma_0 + \theta\log\mu_i \tag{12}$$

The estimated weights would then be $y^{-2\hat{\theta}}$.

In fact, variance function estimation is challenging. Outliers strongly affect (Baumann and Wätzig, 1995) the estimation of variance. Models based on the addition of variance from independent sources are closer to physical reality than the ones based on the contribution of standard deviations.

VARIATION OF PRECISION WITH CONCENTRATION

The precision of an analytical system may be given in function of the analyte concentration (ISO 5725, 1994; Thompson, 1988; Thompson and Howarth, 1973), varying models being proposed as compiled in Table 5.

Table 4. Variant function estimation (Asuero and González, 2007)

	Var (Y_{ij})	w_i	Comments
I	$\sigma_0^2\mu_i^2$	μ_i^{-2}	Constant CV equal to σ; reasonable approach in e.g., HPLC, as long as the limits of assay sensitivity are not approached too closely
II	$\sigma_0^2\mu_i$	μ_i^{-1}	Quite useful from count data for which a Poisson assumption implies that Var (Y_{ij}) = μ_i
III	$\sigma_0^2(\alpha+\mu_i)^2$	$(\alpha+\mu_i)^{-2}$	α known
IV	$\sigma_0^2\mu_i^{2\theta}$	$\mu_i^{-2\theta}$	General model to accommodate overdispersion; θ often falls in the range 0. 6 ≤ θ ≤ 0. 9. Poisson model if $\sigma_0 = 1$ and θ = 0. 5. Power of the mean variance function, which is likely to be of the most importance in chromatographic and capillary electrophoresis applications. Plot of log $\lvert r_{ij}\rvert$ versus log of the predicted value gives a straight line
V	$\sigma_0^2(\theta_1+\mu_i^{\theta_2})$	$(\theta_1+\mu_i^{\theta_2})^{-1}$	θ1 describes the imprecision of measurement that dominates at small response value and θ2 the relationship between mean and variance that dominates at larger response values
VI	$\sigma_0^2\exp(2\theta\mu_i)$		The variability increase very quickly with the mean Plot of log $\lvert r_{ij}\rvert$ versus predicted value $\widehat{\square}_{\square}$ show a linear relationship
VII	$\sigma_0^2(1+\theta_1\mu_i+\theta_2\mu_i^2)^2$		The Standard deviation is thought to be a quadratic function of the dependent variable Plot of log $\lvert r_{ij}\rvert$ versus y_i shows a quadratic relationship
VIII	$\sigma_0^2(1+\theta_1\mu_i+\theta_2\mu_i^2)$		
IX	$\sigma_0^2 g^2(\mu_i,z_i,\theta)$	$[g^2(\mu_i,z_i,\theta)]^{-1}$	g is the variance function general model

Table 5. Varying models

Model	Function	Authors
I	$s = k\sqrt{c}$	ISO 5725 Hughes and Hurley
II	$s_c = s_0 + kc$ $\sigma_c = pc + q$	Thompson and Howarth, 1973 Thompson, 1976 Howarth and Thompson, 1976 Thompson and Howarth, 1978 Thompson, 1978 Thompson, 1988 Lee and Ramsey, 2001
III	$s_x = a_0 + a_1 x + a_2 x^2$	Oppenheimer et al., 1983 Watters et al., 1987 Zorn et al., 1996
IV	$s_c = A_0 + A_1 c + A_2 c^2 + A_3 c^3$	Modamio et al., 1996
V	$s_x = \sqrt{a_0 + a_1 x}$	ISO 11843-2, 2000
VI	$s_x = \sqrt{a_0 + a_1 x^2}$ $s_x = \sqrt{s_0^2 + k^2 c^2}$ $\sigma_x = \sqrt{p^2 c^2 + q^2}$ $u_x = \sqrt{s_0^2 + (x s_i)^2}$	Zitter and God, 1971 Thompson, 1988 Rocke and Lorenzato, 1995 Lee and Ramsay, 2001 Rocke et al., 2003 Wilson et al., 2004 EURACHEM/CITAT Guide, 2002 Heydorn and Anglow, 2002
VII	$\sigma_x = \sqrt{c_0 + c_1 x + c_2 x^2}$ $\sigma_y = \sqrt{c_0 + c_1 y + c_2 y^2}$	Watters et al., 1987 Schwartz, 1978 Boumans et al., 1981 Bubert and Klockenk¨amper, 1983 Oppenheimer et al., 1983
VIII	$\sigma = bc^d$ $s_x = a_0 e^{a_1 x}$ $\sigma_c = pc^k + q$ $\sigma_y^2 = Ay^b$ $\sigma_x^2 = A(x+1)^b$	ISO 5725 Hughes and Hurtley, 1987 Zorn et al., 1997 Desimoni, 1999 Prudnikov and Shapkings, 1984 Oppenheimer et al., 1983

The lineal model, II, is the most simple. When the analytical errors stem from two independent terms, the most satisfactory option should be to combine variances. Then the models V and VI describe the variation of precision with concentration more correctly. Standard deviation usually increases with concentration, whereas the relative standard deviation (coefficient of variation) remains constant or slightly decreases. Some empirical models such as III and VIII have found use for radioassay ligand and other general situations; the standard deviation is modelled as a function of concentration.

The topic concerning the weighting choice is an open subject, and there is no universal solution to this problem being (Modamio et al., 1996) often subjective and somewhat arbitrary.

APPLICATIONS

Some examples of application of heterogeneous variance (weighted least squares) in analytical chemistry are compiled in Table 6. In addition, an experimental situation covering the area of enzymatic kinetics has been subject of study in this book chapter in order to weighting or not properly the data.

Table 6. Some selected applications of weighted least squares in analytical chemistry

Content	Reference
Theory of chromatographic detection and modern approaches to data acquisition and processing is given in the context of the calibration problem	Asnin, 2016
Characterizing nonconstant instrumental variance in emerging miniaturized analytical techniques	Noblitt et al., 2016
Simultaneous determination of 40 novel psychoactive stimulants in urine by liquid chromatography–high resolution mass spectrometry and library matching	Concheiro et al., 2015
Practical guidelines for reporting results in single- and multi-component analytical calibration	Olivieri, 2015

Content	Reference
Method validation using weighted linear regression models for quantification of UV filters in water samples	Pereira da Silva et al., 2015
Using Least Squares for Error Propagation. Practical examples.	Tellinghuisen, 2015
Analysis and interpretation of enzyme kinetic data	Cornish-Bowden, 2014
Selecting the correct weighting factors for linear and quadratic calibration curves with least-squares regression algorithm in bioanalytical LC-MS/MS assays and impacts of using incorrect weighting factors on curve stability, data quality, and assay performance	Gu et al., 2014
Impact of calibrator concentrations and their distribution on accuracy of quadratic regression for liquid chromatography–mass spectrometry bioanalysis	Tan et al., 2014
Reducing the number of signals needed to perform LW calibrations by developing models of weighing factors robust to daily variations of instrument sensibility: Application to the identification of explosives by ion chromatography	Brasil et al., 2013
Comparative study of some robust statistical methods: weighted, parametric, and nonparametric linear regression of HPLC convoluted peak responses using internal standard method in drug bioavailability studies	Korany et al., 2013
The quality coefficient as performance assessment parameter of straight line calibration curves in relationship with the number of calibration points.	de Beer et al., 2012
The approaches for estimation of limit of detection for ICP-MS trace analysis of arsenic	Rajakovic et al., 2012
A comparison in the evaluation of measurement uncertainty in analytical chemistry testing between the use of quality control data and a regression analysis	Sousa et al., 2012
Application of a special in-house validation procedure for environmental–analytical schemes including a comparison of functions for modelling the repeatability standard deviation	Brüggemann and Wennrich, 2011
Overall calibration procedure via a statistically based matrix-comprehensive approach in the stir bar sorptive extraction–thermal desorption–gas chromatography–mass spectrometry analysis of pesticide residues in fruit-based soft drinks	Lavagnnini et al., 2011
Using R2 to compare least square fit models: when it must fail.	Tellinghuisen y Bolster, 2011
Comparison of three weighting schemes in weighted regression analysis for use in a chemistry laboratory	Jain, 2010
Method validation for the endocrine disruptors and pesticides in water by gas chromatography–tandem mass spectrometry using weighted linear regression schemes	Mansilha et al., 2010

Table 6. (Continued)

Content	Reference
Calibration in atomic spectrometry: A tutorial review dealing with quality criteria, weighting procedures and possible curvatures	Mermet, 2010
Comparison between ordinary least squares regression and weighted regression in the calibration of metals present in human milk	Nascimento et al., 2010
Cochran's test optimized "G test": Expressions are derived to calculate upper limit as well as lower limit critical values for data sets of equal and unequal size at any significance level.	't Lam, 2010
Least-squares analysis of data with uncertainty in x and y: A Monte Carlo methods comparison	Tellinghuisen, 2010a
Least-Squares Analysis of Phosphorus Soil Sorption Data with Weighting from Variance Function Estimation: A Statistical Case for the Freundlich Isotherm	Tellinghuisen, 2010b
Least-squares analysis of phosphorus soil sorption data with weighting from variance function estimation	Tellinghuisen and Bolster, 2010
The guiding role of the assumptions for least-squares regression in practical problem solving: Calibration of 109Cd KXRF systems	Brito et al., 2009
Weighted least-squares regression with different weighting functions: Calibration of 109Cd KXRF systems	Brito and Chettle, 2009
Verifying if alternative approaches are available for getting acceptably approximate estimates of the limit of detection	Desimoni and Brunetti, 2009
Least squares in calibration: weights, nonlinearity, and other nuisances	Tellinghuisen, 2009a
The least-squares analysis of data from binding and enzyme kinetics studies: weights, bias, and confidence intervals in usual and unusual situations	Tellinghuisen, 2009b
Weighting Formulas for the Least-Squares Analysis of Binding Phenomena Data	Tellinghuisen, 2009c
Variance function estimation by replicate analysis and generalized least squares: A Monte Carlo comparison	Tellinghuisen, 2009d
Weighting formulas for the least-squares analysis of binding phenomena data	Tellinghuisen and Bolster, 2009
Analysis of Flavonoids in Oxytropis kansuensis Bunge by RP-LC–DAD with Weighted Least-Squares Linear Regression	Li et al., 2008
Least squares with non-normal data: estimating experimental variance functions	Tellinghuisen, 2008a
The problem with using "quality coefficients" to select weighting formulas	Tellinghuisen, 2008b

Content	Reference
Least-squares variance component estimation. Various examples are given to illustrate the theory	Teunissen and Amiri-Simkooei, 2008
Weighted least squares in calibration: Estimating data variance functions in high-performance liquid chromatography	Zeng et al., 2008
A statistical overview on univariate calibration, inverse regression, and detection limits: application to gas chromatography/mass spectrometry technique	Lavagnini and Magno, 2007
A general approach to heteroscedastic linear regression: The methodology is applied to a number of simulated and real examples	Leslie et al., 2007
Weighted least-squares in calibration: The distinction between a priori and a posteriori parameter standard errors is emphasized	Tellinghuisen, 2007
Why are we weighting? Recommendations	Thompson, 2007
Reviews calibration-, uncertainty-, and recovery-related documents from 10 consensus-based organizations	Vanatta and Coleman, 2007
Determination of lanthanides in international geochemical reference materials by reversed-phase high-performance liquid chromatography using error propagation theory to estimate total analysis uncertainties	Santoyo et al., 2006
Understanding Least Squares through Monte Carlo Calculations	Tellinghuisen, 2005b

Enzymatic Kinetics: Lineweaver-Burk, Hones or Eadie- Hofstee?

An example of a function that can be transformed into the hyperbola:

$$y = \frac{x}{\alpha x - \beta} \tag{13}$$

If we do $Y=1/y$, and $X=1/x$, we have the lineal equation:

$$Y = \alpha - \beta X \tag{14}$$

The Eqn. (13) is important because it is the functional form that explains the relationship between the rate of an enzymatic reaction, v, and the concentration of substrate C. The rate of reaction of the enzymatic reactions depends on the affinity of the substrate and the enzyme. If it follows an ideal model ([substrate] >> [enzyme]), and if:

$$E + S \leftrightarrow ES \leftrightarrow E + P \tag{15}$$

A plot of the concentration of the reaction rate versus the substrate concentration follows a rectangular hyperbola through the origin. The relationship between the reaction rate and the substrate concentration can be expressed (Jurs, 1986; Noggle, 1993) by the Michaelis-Menten equation:

$$v = \frac{C}{C + K_m} V_{\max} \tag{16}$$

where K_m and V_{max} are the constants of Michaelis. V_{max} is the reaction rate when the enzyme is completely saturated with the substrate and the reaction proceeds at the maximum possible speed, and K_m is the substrate concentration at half the maximum speed. The Michaelis-Menten equation can be regrouped to produce different linear forms:

Lineaweaver-Burk (LB); plotting *1/v* versus *1/C*:

$$\frac{1}{v} = \frac{1}{V_{\max}} + \frac{K_m}{C V_{\max}} \tag{17}$$

Hanes (H); plotting *C/v* versus *C*:

$$\frac{C}{v} = \frac{C}{V_{\max}} + \frac{K_m}{V_{\max}} \tag{18}$$

Eadie-Hosfstee (EH); plotting v versus v/C:

$$v = V_{\max} - \frac{vK_m}{C} \tag{19}$$

Table 7. Data of substrate concentration and rates for an enzimatic reaction

C [(a)]	*v*	C [(b)]	*v*	C [(c)]	*v*	C [(d)]	*v*	C [(e)]	*v*
0.10	1.9	20.0	48.24	0.03	0.14	0.14	0.15	1	0.10
0.33	4.2	12.0	40.09	0.04	0.17	0.22	0.17	3	0.24
1.00	6.1	8.0	41.19	0.05	0.18	0.29	0.23	5	0.30
3.33	6.5	6.4	37.39	0.09	0.26	0.56	0.32	8	0.54
10.00	7.2	4.8	32.91	0.13	0.31	0.77	0.39	10	0.72
33.30	7.4	3.2	28.72	0.22	0.35	1.46	0.49	15	0.97
100.00	6.9	1.6	18.94	0.43	0.40			20	1.15
				0.65	0.44			30	1.52
				1.08	0.45			40	1.72
								50	1.97

[(a)] Deshidratación de l-malato catalizada por fumarasa (Noggle, 1993)

[(b)] Hipurato de metilo-quimiotripsina a pH 7.8 y 25 ºC (Elmore et al., 1963)

[(c)] Formación de maltosa a partir de almidón/amilasa (Noggle, 1993)

[(d)] Nicotinamida-adenina dinucleótido (Jurs, 1986)

[(e)] (Crabbe, 1982)

The experimental data for various enzymatic systems is shown in Table 7, and the results of the application of e linearization methods unweighted and weighted are shown in Table 8. The application of the least squares in the EH case is questionable. Some of the representations are shown in Figure 1. The weight factors for the LB and H methods are contemplated in Eqn. 20a, b.

$$w_i(LB) = \frac{1}{\left(\frac{\partial(1/v)}{v}\right)^2} = v^4 \qquad w_i(H) = \frac{1}{\left(\frac{\partial(c/v)}{v}\right)^2} = \frac{v^4}{c^2} \tag{20a, b}$$

Both, LB and H methods, lead to identical results by applying the appropriate weights. Transformations distort data space and change the way in which observations affect parameters (Noggle, 1993). There are two ways to avoid these distortions: WLR and nonlinear regression (NLR). Both methods lead to almost similar results.

Table 8. Michaelis constants values obtained by the three methods described by simple linear regression (upper result) and weighted (lower result)

Method	Lineaweaver-Burk		Hanes		Eadie-Hosfstee	
System	K_m	V_{max}	K_m	V_{max}	K_m	V_{max}
I (*)	0.248	7.435	-0.164	6.933	0.268	7.303
	0.242	7.257	0.242	7.257		
II	2.929	5.389	3.036	5.436	2.862	5.352
	2.918	5.409	2.918	5.409		
III (**)	0.073	0.470	0.077	0.478	0.075	0.475
	0.076	0.477	0.076	0.477		
IV	0.441	0.585	0.582	0.685	0.490	0.626
	0.571	0.680	0.571	0.680		
V	16.642	1.766	42.497	3.586	30.271	2.853
	42.253	3.610	42.196	3.806		
K_m	a_1/a_0		a_0/a_1		$-a_1$	
V_{max}	$1/a_0$		$1/a_1$		a_0	

I: L-malato/fumarasa (Noggle, 1993)

II: Methyl Hipurato/quimiotripsina (Elmore et al., 1963)

III: maltosa/amilasa (Noggle, 1993)

IV: nicotinamide/adenina dinucleótid (Jurs, 1986)

V: Crabbe, 1982

NLR (Noggle, 1993):

(*) 0.245±0.068 y 7.24 ± 0.33;

(**) 0.477±0.009 y 0.076±0.05.

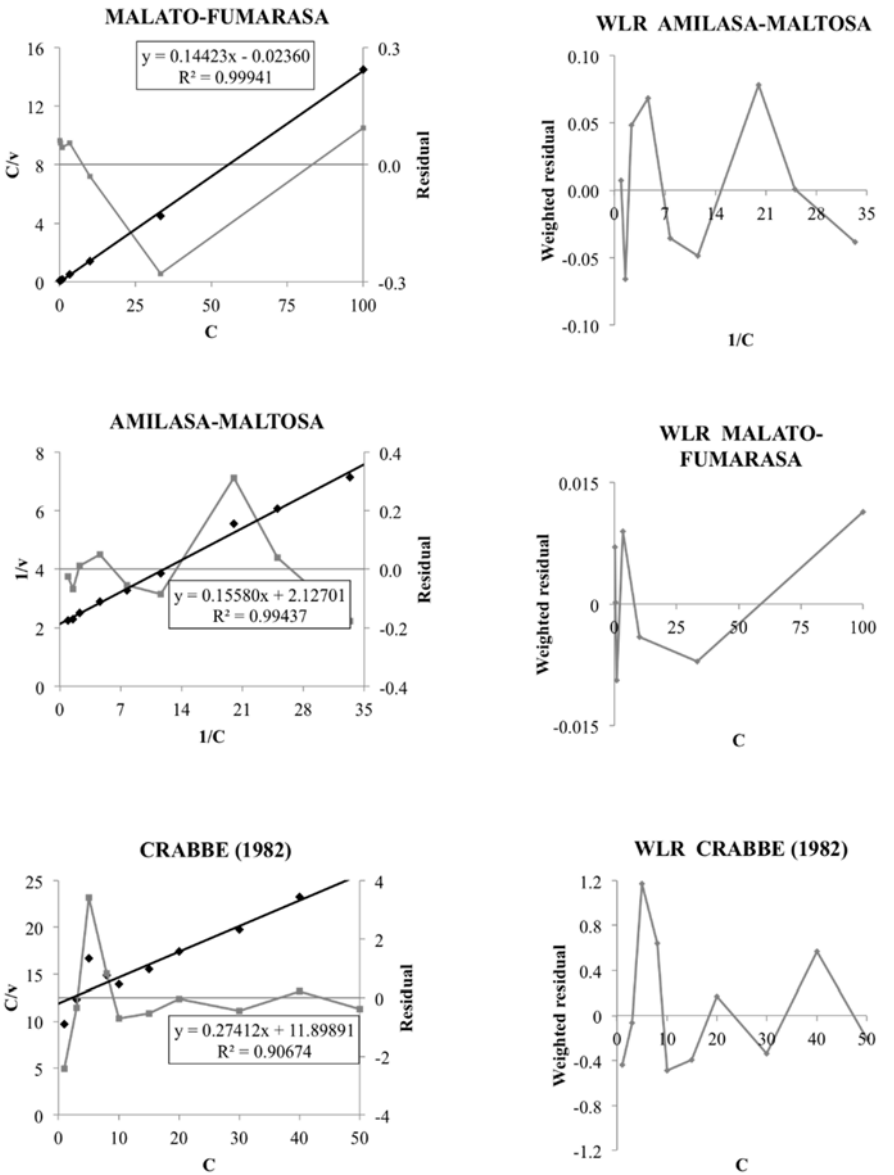

Figure 1. Plots of de Hones and Lineaweaver-Burk for some studies systems.

TRANSFORMING DATA

It may seems that the best way of calculating the coefficient of a non linear equation is the direct application of a non linear regression program. However, NLR is no free (Mager, 1991) from problems: i) depending on the structure of the data and the starting value one may obtained different final solutions; ii) the discrimination between rival models is difficult; iii) NLR is relatively sensitive to deviations from homocedasticity; iv) a substantially multi-collinearity may appears to lead to non robust estimates. Some trouble may be originated using asymptotic NLR estimates (Mager, 1991) because of the too small number of observations in real experiments.

Some advantages are derived from the application of mathematical transformation to experimental data. Transformation may be successfully applied to reach homocedasticity (stabilize variances), to get (an approximate) normality or test in an approximate way the type of model (Meloun and Militký, 2011; Meloun, 1992; Draper and Smith, 1998; Weisberg, 2005).

Graphical or numerical examination of data (Lavagnini and Magno, 2007; Barnet, 2004) may be carried out in order to check (separately or jointly) key assumptions such as linearity of relationship, error independence, residual variance constancy, normally distributed errors, and outliers (Weisberg 2005; Belloto and Sokolovski, 1985). Informal plots may reveal in a clear way the need for a given transformation such as *ln x* or *1/y*, holding in reserve the checking with a more formal analysis (Draper and Smith, 1998). The log rule and the range rule are two often-helpful empirical rules (Weisberg, 2005). Logarithm rule applies when the variable is strictly positive and range over more than order of magnitude. If the range is less than one order of magnitude any transformation is useless. The greater the quotient y_{max}/y_{min} the greater the effect of the transformation considered. As more α differs from the unity greater is the effect of a power transformation of the kind $Y=y^{\alpha}$ (Box and Draper, 1987). Logarithms and exponentials are involved in the most common transformations (EPA, 2000; Daniel and Wood, 1999; Tomassone et al.,

1983). Transformations less common include reciprocals, square roots, and trigonometric functions. Sometimes, two or more of these functions may be combined, as occur (Bysouth and Tyson, 1986) with some calibration programs of commercial instruments.

Lacking prior information trial and error is involved to ascertain the kind of transformation to be used. Plotting on the normal probability paper the accumulative frequency of the transformed variable allows selecting as most suitable the transformation yielding the best straight line. Statistical tests such as comparisons, confidence limits, etc., are thus applied to the transformed data, and results obtained back transformed to the original curvilinear scale (Meloun et al., 1992; Acton, 1959), if desired. Finally, if the transformations are not adequate, proceed to reject the linear model considering instead a series of alternative nonlinear models.

TRANSFORMATIONS TO SIMPLIFY RELATIONSHIPS

Two situations are mainly considered in this context (Asuero and Bueno, 2011; Draper and Smith, 1998; Rawlings et al., 1998). If no priori knowledge of the model fitting the data is available an empirical form of the dependence between the variables is searched so that a straight line relationships may be used. The model is linear in the parameters (Lavagnini and Magno, 2007); and it is only considered the form in which the variables are expressed.

The aim of a transformation is the re-expression, when possible, of a non-linear model into another one linear to which ordinary linear regression may be applied. The functional form of the model dictates the possible transformation.

Simple power transformations suitable only for positive values do not retain the scale and is not always continuous, carrying out transformations for symmetry (Gad, 1999). The power family of transformations $x^* = x^k$ or $y^* = y^k$, named as the "one bend" transformations (Rawlings et al., 1998), affords a useful set of transformations for "straightening" a single bend in the relationship between two variables. The corresponding transformations

may be ordered according to the k exponent values thus obtaining a sequence of power transformations known as (Mosteler and Tukey, 1977; Tukey, 1977) ladder of reexpressions (Table 9).

The sharpness of the curvature determines how far has to move on the transformations ladder (Rawlings et al., 1998). Several transformations on a few observations covering the all range of the data are proved (one independent variable being involved), choosing the transformation that reaches the highest linearity. If the variable follows a Poisson distribution (i.e., count, frequency), the square root transformation is used (less dramatic than taking logarithms) (Shumway et al., 2002). The reciprocal transformation reverses the order of observations having a much more drastic effect than taking logs, being useful when data have an extremely skewed distribution. However, the use of the reciprocal transformation is not common, and the log transformation is preferably to any other if yields satisfactory results.

Table 9. Tukey's Ladder of Transformations
(for choosing a function to change a distributions' shape)
(Asuero and Martín Bueno, 2011)

Need to Correct	Strength of Transformation	Mathematical Function	k exponent value $y^* = y^k$
Positive Skew	Stronger	$-\frac{1}{y^2}$	-2
	Mild	$-\frac{1}{y}$	-1
	"	$\ln y$	0
	"	$\sqrt{y}$	$\frac{1}{2}$
No Shape Change	---	y	1
Negative Skew	Mild	y^2	2
	"	y^3	3
	Stronger	$\exp y$	

TRANSFORMATIONS TO LINEARIZE THE MODEL

Adequately transformations allow non-linear models to be presented in a linear form (Table 10) (de Levie, 2004; Bayne and Rubin, 1986; du Toit et al., 1986; Tomassone et al., 1983; Daniel and Wood, 1980). It should be noted that using of these transformations are certain to accomplish one thing only, i.e., to yield a straight-line form. Some assumptions theoretically necessary to apply the least squares method may not be necessarily satisfied by the transformed data. (Bates and Watts, 2007; Seber, 2003; Belloto and Sokolovski, 1985).

Examples of linearizable functions are shown in Figure 2. The power family of transformations cannot straighten relationships showing more than one bend (i.e., the classical S-shaped growth curve). Logit, arcsin (or angular), and probit are commonly used two-bend transformations kind.

Table 10. Nonlinear function that can be written in a linear form by means of a transformation (Asuero and Martín Bueno, 2011)

Function	Formula	Transformation	Linear form
Power function	$y = \alpha x^b$	$y' = \log y$ $x' = \log x$	$y' = \log \alpha + \beta x'$
Exponential grown model	$y = \alpha e^{\beta x}$	$y' = \log y$	$y' = \log \alpha + \beta x$
Logarithmic	$y = \alpha + \beta \log x$	$x' = \log x$	$y = \alpha + \beta x'$
Hyperbolic	$y = \frac{x}{\alpha x - \beta}$	$y' = 1/y$ $x' = 1/x$	$y' = \alpha - \beta x'$
Logit	$y = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$	$y' = \log\left(\frac{y}{1-y}\right) =$ $2\tanh^{-1}(2y-1)$	$y' = \alpha + \beta x$

In transforming data, the disturbance term is also transformed, thus affecting the assumptions concerning it (Bates and Watts, 2007; Seber, 2003; Box and Draper, 1987). That is to say, correct assumptions of

additive and normal disturbance terms for model functions are not valid for the transformed data. Use non-linear regression on the original data, or weighted least squares on the transformed data being then required. Fitting then the transformed model leads to some initial estimates (Meloun et al., 1992; Mager, 1991).

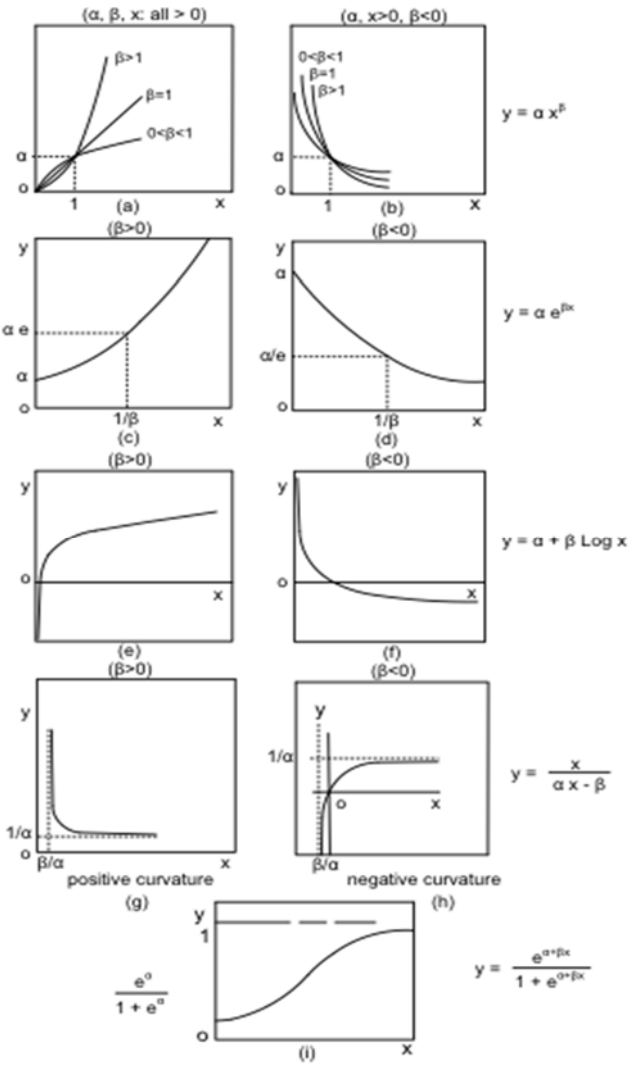

Figure 2. Plots of linearizable curves (Asuero and Martín Bueno, 2011).

TRANSFORMATIONS TO STABILIZE VARIANCES

Non constant variance are related with non normal distributed data (Canavos 1984, Rios 1977), being data transformation the most appropriate mean to deal with such situations (Asuero and Martín Bueno, 2011). Variance heterogeneity usually appears when the errors corresponding to some treatments are significantly higher (or lower) than others, given the nature of the experiences. In a normal distribution, the variance σ_y and the mean σ_y are independent; a direct relationship between the mean and the variance is typical from all other common distributions. Either theoretical considerations and/or a preliminary empirical analysis may suggest the nature of the dependence between the variance and the mean value (Box and Draper, 1987). If the functional relationship is known, a transformation exists making (approximately) constant the variance (Draper and Smith, 1998). With certain kinds of data, heterogeneous (non uniform) variance and non normality are expected at first. The same experimental situations that lead to non-normal distributions as usually provide heterogeneous variances as $\sigma_y = f(\bar{\square})$ in most non- normal distributions (Brownlee, 1984; Natrella, 1963).

Table 11 summarizes a number of transformations (some from the power family) used to correct for homogeneity and approximate normality. Note that in stabilizing variance, the transformed variable is more normal (Gaussian).

TRANSFORMATION BASED ON SAMPLE DATA OBSERVATIONS: BOX-COX METHOD

There are many situations in which the only information available in order to determine the appropriate transformation is the sample data. Residual plots may indicate a given kind of transformation to be applied. Several transformations may be applied and it should be adopted the one

that better meet the normality criteria. The most appropriate transformation can also be calculated empirically (Box and Cox, 1964).

Table 11. Transformation to correct for homogeneity and approximate normality (Asuero and Martín Bueno, 2011)

Data type	$s_y = f(y)$	Variance stabilizing transformation
Poisson (Count)* $y \geq 0$	$\sqrt{y}$	$\sqrt{y}$
Small counts (y)**		$\sqrt{y+1}$ or $\sqrt{y}+\sqrt{y+1}$
Binomial ($0 < y < 1$)	$\sqrt{y(1-y)}$	$a\sin\sqrt{y}$
Negative binomial $0 \leq y \leq 1$	$\dfrac{\sqrt{1-y}}{y}$	$\left(\dfrac{2+y}{3}\right)\sqrt{1-y} = (1-y)^{1/2} - \dfrac{(1-y)^{3/2}}{3}$
Variance = (mean)^2 $y \geq 0$	y	$\ln y$
Correlation coefficient $-1 \leq y \leq 1$	$\dfrac{1}{\sqrt{1-y^2}}$	$0.5\left[\ln(1+y) - \ln(1-y)\right]$

* Modifications for the Poisson and binomial cases have been suggested by Freeman and Tukey (105).

** It should be noted that the square root transformation overcorrects when very small values and zero appears in the original data. In these cases, $\sqrt{y+1}$ is often used as a transformation.

A collection of transformations characterized by one or a few selected parameters (Weisberg, 2005; Chinn, 1996; Sakia, 1992) is called a family of transformations.

Logarithmic, square root, and inverse transformations are contained in the Tukey (1957) family (Table 12):

$$T = \begin{cases} y^{\lambda}, & for\ \lambda \neq 0 \\ \ln y, & for\ \lambda = 0 \end{cases} \tag{21}$$

$\lambda \leq 1$. When $\lambda \to 0$ then $y^{\lambda} \to 1$ (discontinuity at λ zero values), and the transforation lacs of sense (Draper and Smith, 1998). In order to choose the best λ value to run smoothly as λ approaches zero, Box and Cox (1964) carried out the following proposal (Chinn 1996; Sakia, 1992; Peace, 1988; Schlesselman, 1971)

$$W = \begin{cases} (y^{\lambda} - 1)/\lambda, & for \lambda \neq 0 \\ \ln y, & for \lambda = 0 \end{cases} \tag{22}$$

identically in essence to Eqn. 21 for those cases in which a constant term b0 is contained in the regression model the (Peace, 1988).

The same transformation was practically suggested by Kapteyn (1916; 1903) about 150 years ago in his work on growth. Sclove (1972) gave a test of λ=0 (*log y* versus *x*). As the *F* statistic in the analysis of variance is invariant under linear transformations (Malaeb, 1997), Eqns. 21 and 22 are equivalent.

It is more convenient, however, to employ (Table 12)

$$V = \begin{cases} (y^{\lambda} - 1)/(\lambda \dot{y}^{\lambda - 1}), & for \lambda \neq 0 \\ \dot{y} \ln y, & for \lambda = 0 \end{cases} \tag{23}$$

being $\dot{y}$ the geometric mean of the corresponding y_i values

$$\dot{y} = (y_1 y_2 \ldots y_k)^{1/k} = (\prod y_i)^{1/k} \tag{24}$$

$\dot{y}$ is a constant, evaluated at the beginning of the calculations by means of the formula

$$\ln \dot{y} = \frac{\sum_{i=1}^{k} \ln \dot{y}_i}{k} \tag{25}$$

As the values of λ varies, the *W's* may change drastically (disadvantage of Eqn. 22), posing minor problems; a special program being required to achieve the best λ value. It is then better to use Eqn. (23).

The $\dot{y}^{\lambda-1}$ in Eq. (23) is the *nth* power of the appropriate Jacobian of the transformation (Mateu, 1997); the set of y_i values are converted into the W_i ones

$$J(\lambda, y) = \prod_{i=1}^{k} \left| \frac{dy_i^{(\lambda)}}{dy_i} \right| = \prod_{i=1}^{k} y_i^{\lambda-1} \tag{26}$$

A more flexible extended power or shifted power transformation family was also proposed in the Box and Cox (1964) paper, which accounts for negative y's

$$y(\lambda) = \begin{cases} (y+\lambda_2)^{\lambda_1} - 1 \Big/ \lambda_1 & if\ \lambda_1 \neq 0 \\ \log(y+\lambda_2) & if\ \lambda_1 = 0 \end{cases} \tag{27}$$

In practice we get $y + \lambda_2 > 0$ for any y value. The variance of the untransformed scale must be proportional to $(mean + \lambda_2)^{(1-\lambda_1)}$ (Chinn, 1996), when we deal with a shifted variance stabilizing Box-Cox transformation. Approximate normality is only to be expected (Sakia, 1992) because the $y^{(\lambda)}$ range (Eqns. 22, 23 and 27) is restricted.

It is important to note that the range of equations (22)-(23) and (27) is restricted according to whether λ is positive or negative. This implies that transformed values do not cover the entire range $(-\infty, +\infty)$ (bounded supported distribution).

The idea behind the Box and Cox method (Draper and Smith, 1998) is that, if a suitable λ value is found, then it is possible by the maximum

likelihood method fitting an additive model with a normal, independent and homogeneous error structure (Draper and Smith, 1998).

Box-Cox transformation calculations are carried out by means of computer programs (Huang et al., 1978; Chang, 1977). Statistical packages including graphical facilities allow selecting power transformations (Weisberg, 2005; Malaeb, 1997) of both the independent and dependent variables. Multivariate Box-Cox transformation (Rode and Chinchilli, 1988) may be performed by means of the MULTTBXX program (the univariate transformation being a special case).

Table 12. Values of certain power functions for transformations to stabilize variances (Asuero and Martín Bueno, 2011)

λ	y^{λ}	$W=\left(y^{\lambda}-1\right)/\lambda$	$V=\left(y^{\lambda}-1\right)/\left(\lambda\,\dot{y}^{\lambda-1}\right)$
1	y	$y-1$	$y-1$
0.5	$\sqrt{y}$	$2\left(\sqrt{y}-1\right)$	$2\sqrt{\dot{y}}\left(\sqrt{y}-1\right)$
0	1(?)	$\ln y$	$\dot{y}\ \ln y$
-0.5	$1/\sqrt{y}$	$2\left(1-1/\sqrt{y}\right)$	$2\dot{y}\sqrt{\dot{y}}\left(1-1/\sqrt{y}\right)$
-1	y^{-1}	$2\left(1-1/\sqrt{y}\right)$	$\dot{y}^{2}\left(1-1/\sqrt{y}\right)$

APPLICATIONS

Some examples of application of transforming data and Box-Cox transformation in analytical chemistry are compiled in Tables 13 and 14.

An illustrative example taken form scientific journals about the equilibrium constant for the heterogeneous reaction and the CO_2 vapour pressure versus temperature is selected in order to elucidate this book chapter:

Equilibrium Constant for the Heterogeneous Reaction: $CoTiO_3 = Co + TiO_2 + CO_2$

The dependence of the equilibrium constant on temperature has the form:

$$\log K = \frac{\Delta S}{4.576} - \frac{\Delta H}{4.576} \cdot \frac{1}{T} \tag{28}$$

ΔS and ΔH are the entropy and enthalpy variations of the reaction, expression of the type

$$Y = a_0 + a_1 x \tag{29}$$

$$Y = \log K \qquad a_0 = \frac{\Delta S}{4.576} \qquad a_1 = -\frac{\Delta H}{4.576} \qquad x = \frac{1}{T} \tag{30}$$

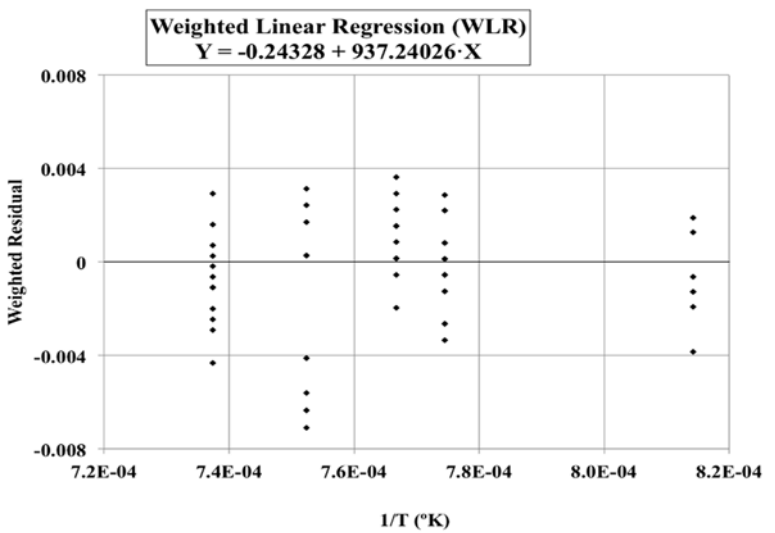

Figure 3. Plot of the measurement of the equilibrium constant of the reaction as a function of the temperature. Middle: Plot of the transformed data (*log* K_{eq}) in function of *1/T* ºK (Single Linear Regression). Bottom: Residual analysis.

The results obtained on the measurement, at five different temperatures, of the equilibrium constant for the reaction $CoTiO_3$ = Co + TiO_2 + CO_2 are shown in Table 15, and in Figure 3 (top). Also, in Figure 3 (center) is plotted the straight line that adjusts to the transformed data, *log* K_{eq} in function of *1/T* ºK, together with the residual analysis (Figure 3 bottom) by simple linear regression.

Table 13. Some selected applications of transformations

Content	Reference
A bilogarithmic hyperbolic cosine method for the evaluation of overlapping formation constants at varying (or fixed) ionic strength	Beaumount et al., 2016
Evaluation of three isotherm models (Langmuir, Freundlich, and Dubinin-Radushkevich) to correlate four sets of experimental adsorption isotherm data, which were obtained by batch tests in lab	Chen, 2015
Kinetics of Carbaryl Hydrolysis: An Undergraduate Environmental Chemistry Laboratory	Hawker, 2015
Feasibility study of potentiometric multisensor system of 18 ion-selective and cross-sensitive sensors as an analytical tool for determination of urine ionic composition	Yaroshenko et al., 2015
A novel multiple headspace extraction gas chromatographic method for measuring the diffusion coefficient of methanol in water and in olive oil	Zhang and Chai, 2015
Adsorption Kinetics and Isotherms: A Safe, Simple, and Inexpensive Experiment for Three Levels of Students	Piergiovanni, 2014
Evaluation of Equilibrium Sorption Isotherm Equations: Datasets from literatures are selected and three two-parameter and three-parameter equations were used to evaluate adsorption systems	Chen, 2013
Statistical Analysis of Linear and Non-linear Regression for the Estimation of Adsorption Isotherm Parameters	Osmari et al., 2013
Equlibrium sorption of the phosphoric acid modified rice husk: Langmuir, Freundlich, Temkin and Dubinin–Radushkevich Isotherms Studies	Dada et al., 2012
Application of the van't Hoff dependences in the characterization of molecularly imprinted polymers for some phenolic acids: Evaluation of the temperature effect on the sorption processes investigated analytes in methanol and acetonitrile (porogen) as mobile phases	Denderz and Lehotay, 2012
An alternative analytical method for measuring the kinetic parameters of the enzymes invertase and lactase	Heinzerling et al., 2012

Table 13. (Continued)

Content	Reference
Study of the pattern of formation of absorption signals for high concentrations of analyte atoms in the absorption volume and to employ the findings for High-resolution continuum source electrothermal atomic absorption spectrometry data quantification within a broad concentration range of the analyte	Katskov et al., 2012
A comprehensive treatment of experimental enzyme kinetics strongly coupled to electronic data acquisition and use of spreadsheets to organize data and perform linear and nonlinear least-squares analyses	Barton, 2011
Chemical Dosing and First-Order Kinetics: Examples of multiple-dose problems are presented that are appropriate for students taking introductory, general, and physical chemistry courses	Hladky, 2011
On the use of linearized pseudo-second-order kinetic equations for modeling adsorption systems	El-Khaiary et al., 2010
Insights into the modeling of adsorption isotherm systems: accuracy and consistency in parameters prediction or estimation	Foo and Hameed,□2010
Introduce and compare numerical approaches that involve diferent levels of knowledge about the noise structure of the analytical method used for initial and equilibrium concentration determination	Markovic et al., 2010
Polydimethylsiloxane-based permeation passive air sampler. Part II: Effect of temperature and humidity on the calibration constants	Seethapathy and Górecki, 2010
A simple competitive enzyme-linked immunosorbent assay (cELISA) was established for rapid measure- ment of secretory immunoglobulin A (sIgA) in saliva	Wang et al., 2010
An equation relating the absorbance of the solute to the acidity constants (pKa1 and pKa2) and pH is derived for weak diprotic acids (diprotic bases and zwitterions)	Asuero, 2009
Weighting Formulas for the Least-Squares Analysis of Binding Phenomena Data	Tellinghuisen, 2009c
A comprehensive study on the possibility of applying the nth-degree polynomial logistic regression model for fitting the kinetic conversion data of cellulose pyrolysis	Cai et al., 2008
The Hill equation: a review of its capabilities in pharmacological modelling	Goutelle et al., 2008
Least-squares regression of adsorption equilibrium data: comparing the options	El-Khaiary, 2008
Evaluation of logistic and polynomial models for calibration curves spanning the quantitative concentration range for seven different protein assays based on examination of residuals	Herman et al., 2008

Content	Reference
Methods for studying reaction kinetics in gas chromatography, exemplified by using the 1-chloro-2,2-dimethylaziridine interconversion reaction	Krupcık et al., 2008
A bilogarithmic hyperbolic cosine method for the spectrophotometric evaluation of stability con- stants of 1: 1 weak complexes is developed and applied to data found in the literature	Boccio et al., 2007
Examination of the limitations of using linearized Langmuir equations by fitting P sorption data collected on eight different soils with four linearized versions of the Langmuir equation and comparing goodness-of-fit measures and fitted parameter values with those obtained with the nonlinear Langmuir equation.	Bolster and Hornberge, 2007
A review of existence criteria for parameter estimation of the Michaelis–Menten regression model	Jukic et al., 2007
Highlights some common errors of data evaluation that are frequently found in the literature	Badertscher and Pretsch, 2006
The general equation resulting from the logistic transformation is discussed considering the stoichiometric factors for monovalent anions, and the linearization of the theoretical fit to experimental data was checked for two real cases	Capitán-Vallvey et al., 2006
Alternative method to the Arrhenius equation for termogravimetric analysis based on a logistic mixture model	Naya et al., 2006
Log-log transformation without weighting is the simplest model to fit the calibration data for the determination of piperaquine (PC) in urine	Singtoroj et al., 2006
A bilogarithmic hyperbolic cosine method for the spectrophotometric evaluation of stability constants of 1:1 weak complexes from continuous variation data is devised and applied to literature data	Sayago and Asuero,2006

Table 14. Some selected papers on Box-Cox transformation

Content	Reference
Estimating Box-Cox power transformation parameter via goodness of fit tests. An artificial covariate method is also included for comparative purposes	Asar et al., 2017
Two strategies are proposed to extend and unify residual error modeling: a dynamic transform-both-sides approach combined with a power error model capable of handling skewed and/or heteroscedastic residuals, and a t-distributed residual error model allowing for symmetric heavy tails	Dosne et al., 2016
Models with Transformed Variables.Interpretation and Software	Boef et al., 2015

Table 14. (Continued)

Content	Reference
Overview of state-of-the-art dose-response analysis, both in terms of general concepts that have evolved and matured over the years and by means of concrete examples	Ritz et al., 2015
Optimization of sonochemical degradation of tetracycline in aqueous solution using a central composite design	Safari et al., 2015
New methodology for estimating λ and an alternative method of determining plausible values for it	Vélez et al., 2015
Experimental design and multiple response optimization. Using the desirability function in analytical methods development	Candioti et al., 2014
Design-based development of a stability-indicating RP-HPLC method for the simultaneous determination of parabens in pharmaceutical formulation	Roy and Chakrabarty, 2014
Occurrence of pharmaceuticals in urban wastewater of north Indian cities and risk assessment	Singh et al., 2014
Statistical Evaluation and Validation of Quantitative Methods of Drug Analysis	Komsta, 2013
A calibration-free/minimum approach, iterative optimization technology, which is used to predict (without calibration standards) the composition of a mixture while maintaining a similar predictability to calibration standard models	Muteki et al., 2013
Gaussian Quadrature is an efficient method for the back-transformation□in estimating the usual intake distribution when assessing dietary exposure	Dekkers and Slob, 2012
CALUX measurements: Statistical inferences for the dose–response curve. Use of linear calibration functions based on Box–Cox transformations to overcome the issue of uncertainty assessment	Elskens et al., 2011
Statistical Data Analysis in data transformation: A practical guide	Meloun and Militký, 2011
A general equation is presented for modeling retention, using the organic modifier content of the mobile phase. The equation is based on the Box-Cox transform of modifier concentration.	Komsta, 2010
Overview of traditional normalizing transformations and how Box-Cox incorporates, extends, and improves on these traditional approaches to normalizing data. Examples of applications are presented, and details of how to automate and use this technique are included	Osborne, 2010
Least-Squares Analysis of Phosphorus Soil Sorption Data with Weighting from Variance Function Estimation: A Statistical Case for the Freundlich Isotherm	Tellinghuisen, 2010

Content	Reference
Evaluation of the environmental contamination at an abandoned mining site using multivariate statistical techniques. The Box–Cox transformation has been used to transform the data set in normal form in order to minimize the non-normal distribution of the geochemical data	Bagur et al., 2009
A method for identifying relevant proteins from SIMCA discriminating powers is proposed, based on the Box-Cox transformation coupled to probability papers	Marengo et al., 2008
Application of differential permeation and Box–Cox transformation in the analysis of di-n-octyl disulfide in a straight oil metalworking fluid	Xu and Que Hee, 2006
The Box-Cox transformation applied to soil data improves sample symmetry and stabilizes spread; the logarithmic plot of a profile likelihood function enables the optimum transformation parameter to be found□	Meloun et al., 2005

Table 15. Equilibrium constant for the heterogeneous reaction $CoTiO_3 = Co + TiO_2 + CO_2$ (Spiridonov y Lopatkin, 1973)

T ºK	955	1018	1031	1056	1083
K_{eq}	3.30	3.00	2.96	2.81	2.82
	3.25	2.99	3.04	2.81	2.84
	3.29	3.00	2.99	2.84	2.78
	3.29	3.04	3.02	2.80	2.75
	3.34	3.02	3.01	2.82	2.80
	3.28	3.04	2.99	2.84	2.74
	3.33	3.03	2.99	2.94	2.71
		3.03	3.01	2.90	2.76
		3.08	3.00	2.92	2.87
		3.07	3.03	2.93	2.81
		3.07	2.98	2.90	2.78
		3.05	3.01	2.90	2.79
		3.07			2.78
		3.05			
		3.02			
		3.04			
		3.05			

Since the number of measurements for the different x_i is different, the Barlett test is used to estimate the homogeneity (homoscedasticity) of the variances

$$B_{\exp} = \frac{1}{C} \cdot \ln \bar{s}^{-2} \sum_{i=1}^{k} f_i - \sum_{i=1}^{k} (f_i \ln s_i^2) \tag{31}$$

$$C = 1 + \frac{1}{3(k-1)} \left(\sum \frac{1}{f_i} - \frac{1}{\sum f_i} \right) \tag{32}$$

If the equality of the variances is true, the magnitude B_{exp} obeys the Chi-square distribution (χ^2_β) with *k-1* degrees of freedom, if each of $f_i > 2$. If $B_{\exp} < \chi^2_\beta(k-1)$, the hypothesis is accepted. If, on the contrary, $B_{\exp} > \chi^2_\beta(k-1)$ the hypothesis is considered incompatible with the experimental data obtained.

Since $B_{\exp} = 13.3 > \chi^2_{0.05}(4) = 9.49$, the hypothesis of homogeneity of the variances is discarded (Table 16), while for a significance level of 1%, $\chi^2_{0.01}(4) = 13.3$ the hypothesis of homogeneity can not be rejected with certainty. In this case, the measures are treated assuming the worst variant, that is heteroscedasticity (other option is to repeat the experiment), and WLR is performed (Table 17), with

$$s^2_{\bar{y}_i} = \frac{s^2_{y_i}}{n_i} \qquad w_i = \frac{c}{s^2_{\bar{y}_i}} \tag{33}$$

where c ($= s^2_{PE}$) is an arbitrary constant that does not influence on the final results of a_0 and a_1, s_{a0} and s_{a1} (although it does affect the magnitudes of $s_{y/x}$ and cov(a_0, a_1)). If we follow the criterion of Spiridonov and Lopatkin (1973) to make the sum of the weights equal to 1 we have

$$\sum w_{\bar{y}_i} = 1 = c\sum \frac{1}{s^2_{\bar{y}_i}} \qquad c = \frac{1}{\sum s^2_{\bar{y}_i}} \qquad w_{\bar{y}_i} = \frac{1}{s^2_{\bar{y}_i}\sum \frac{1}{s^2_{\bar{y}_i}}} \qquad f = \frac{1}{\sum_i w^2_{\bar{y}_i} / f_i} \tag{34}$$

The weighted residuals (Figure 4) show the form of rectangular band (and not funnel) indicating a random distribution.

Table 16. Bartlett's test for homogeneity of variances

N	7	17	12	12	13
Mean	0.5181	0.4826	0.4775	0.4574	0.4451
SD	4.003E-03	3.794E-03	3.214E-03	7.912E-03	6.647E-03
Variance s2	1.603E-05	1.439E-05	1.033E-05	6.260E-05	4.419E-05
Degrees of freedom (f)	6	16	11	11	12
Sum f					56
f * s2	9.616E-05	2.303E-04	1.136E-04	6.886E-04	5.303E-04
(s2 mean)					2.962E-05
ln (s2 mean) * sum f					-583.9083
ln s2	-11.0413	-11.1489	-11.4806	-9.6787	-10.0271
f * ln s2	-66.2475	-178.3821	-126.2870	-106.4655	-120.3247
Sum (f* ln s2)					-597.7068
1/f	0.1667	0.0625	0.0909	0.0909	0.0833
Sum (1/f)					0.4943
B	**13.7985**	**B = ln s2(mean) * sum fi - sum (fi * ln si2)**			
C	**1.0397**	**C=1+[1/(3(k-1)]*[sum (1/f) - 1/(sum f)]**			
B/C	13.272	Chi2 (0.05, 4)	9.488		

B/C = 13.272 > 9.488: The hypothesis of homogeneity of s2 can not be accepted at 5% level. With a significance level of 1% [Chi2(0.01,4) = 13.277]: The hypothesis can not be rejected with certainty.

Table 17. Equilibrium constant of the reaction as a function of the temperature: *log K_{eq}=a_0+a_1·(1/T)* (WLR)

N	**1/T (ºK)**	**log K_{eq}**	**s^2 (log K_{eq})**	**s^2/N**	**w**
7	8.142E-04	0.5181	1.603E-05	2.290E-06	4.368E+05
17	7.745E-04	0.4826	1.439E-05	8.466E-07	1.181E+06
12	7.668E-04	0.4775	1.033E-05	8.607E-07	1.162E+06
12	7.524E-04	0.4574	6.260E-05	5.217E-06	1.917E+05
13	7.374E-04	0.4451	4.419E-05	3.399E-06	2.942E+05
	3.845E-03	2.3807			3.266E+06
w(norm)	**w(i)*x(i)**	**w(i)*y(i)**	**w(i)*x(i)^2**	**w(i)*y(i)^2**	**w(i)*x(i)*y(i)**
0.1337	1.089E-04	0.06929	8.867E-08	0.03590	5.642E-05
0.3617	2.801E-04	0.17456	2.170E-07	0.08424	1.352E-04
0.3558	2.728E-04	0.16987	2.092E-07	0.08111	1.303E-04
0.0587	4.416E-05	0.02685	3.322E-08	0.01228	2.020E-05
0.0901	6.643E-05	0.04009	4.898E-08	0.01785	2.957E-05
1.0000	7.724E-04	0.48067146	5.970E-07	0.231383438	3.716E-04
S(XX)=	3.809E-10	**a1=**	937.24026	**s(y/x)=**	0.001123407
S(XY)=	3.570E-07	**a0=**	-0.24328	**s(a1)=**	57.5606
S(YY)=	3.384E-04	**R=**	0.99439	**s(a0)=**	0.0445
				cov(a0,a1)=	-2.5592

$$F_{\exp}=\frac{s^2_{LOF}}{s^2_{PE}}=\frac{1.2604\cdot 10^{-6}}{3.0621\cdot 10^{-7}}=4.12 \qquad F_{0.05(3,42)}=2.83 \qquad F_{0.01(3,42)}=4.29 \tag{35}$$

The linearity hypothesis cannot be accepted for β=0.05, nor rejected for β=0.01. Since the non-linearity is unlikely in the studied *T* interval, this possibility is attributed to the insufficient accuracy of the experimental data.

It can be taken into account the weights dependent on the transformation

$$w_t=\frac{1}{\left(\frac{\partial \log K}{\partial K}\right)^2}=\frac{1}{\left(\frac{1}{\ln 10}\frac{\partial \ln K}{\partial K}\right)^2}=\ln 10^2 K^2 \tag{36}$$

which are combined to estimated through replicate measures. In this case:

$$Y = (-0.2276 \pm 0.0365) + (916.8915 \pm 46.8490) \cdot x \qquad R = 9.9961 \quad (37)$$

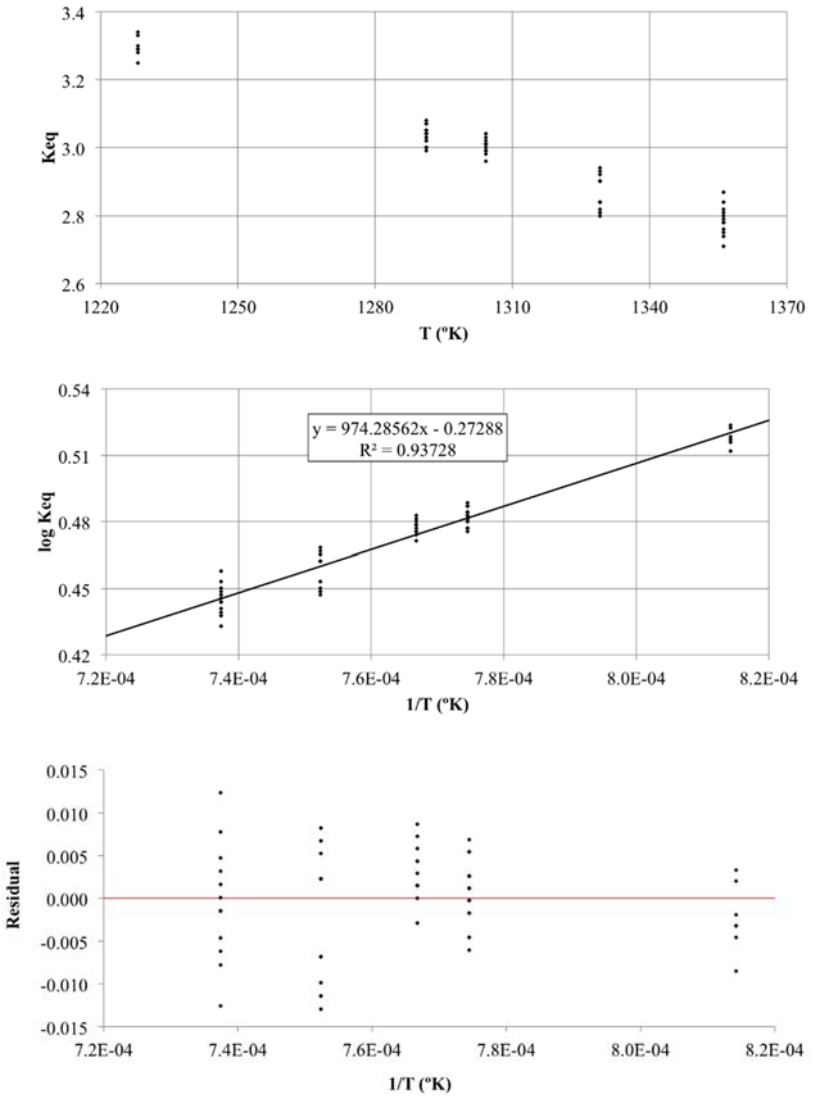

Figure 4. Weighted residuals versus inverse of the temperature; *log* K_{eq}*=f(1/T ºK)*.

The CO_2 Vapour Pressure versus Temperature Case

The linear variation of the physical magnitudes is not universal, although it is often possible to find a coordinate transformation (Asuero and Bueno, 2011) that converts the non linear data into linear ones. The vapour pressure (in atmospheres) of liquid CO_2 as as a function of temperature (in Kelvin degrees) is not linear (Figure 5). There is a theoretical justification (the Claussius-Clapeiron equation) that allows fitting the data of the vapour pressure (P) versus the absolute temperature (K) into an equation of the form:

$$\ln P = A + \frac{B}{T} \tag{38}$$

If we do: $Y = ln\ P$; $X = 1/T$; we have lineal form:

$$Y = A + B \cdot X \tag{39}$$

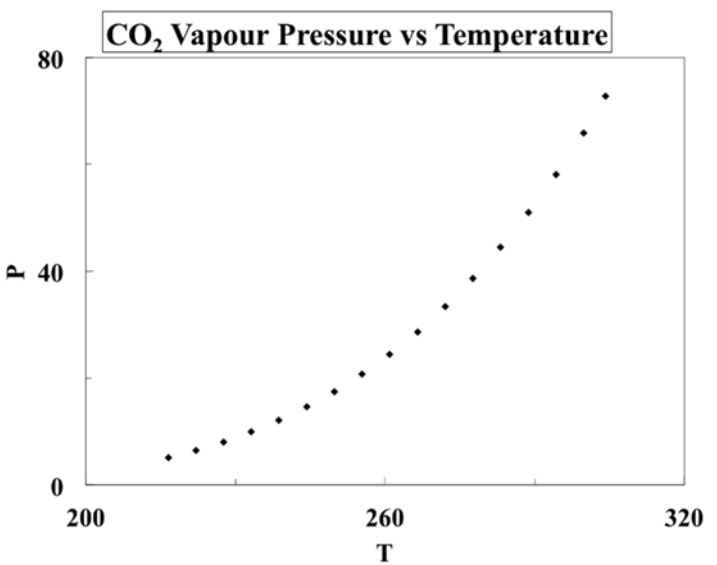

Figure 5. CO_2 vapor pressure data as a function of temperature.

This requires a transformation of the data. After making the appropriate transformations (Table 18), the resulting graph (Figure 6, continue line) is examined. Is the data linear?. Statistics and the graph appear to be fine, and apparently there are no obvious reasons to suspect a problem with this analysis.

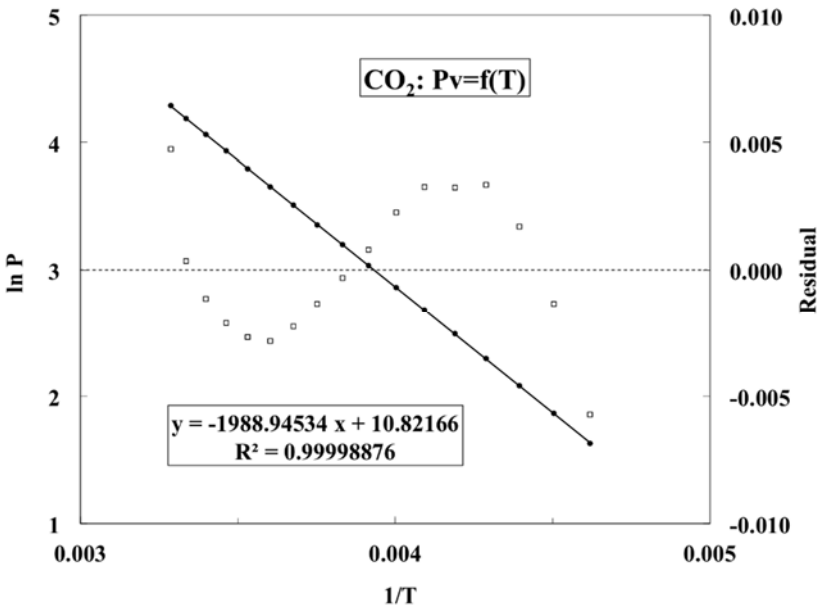

Figure 6. Plot of *ln P* versus *1/T* (Clausius-Clapeiron equation) including residuals analysis.

However, the residuals and the line resulting from the least-squares ($Y = A + BX$) model fitted to the data could be combined in the same plot for checking purposes. The results (Figure 6) lead to a correlation coefficient of 0.99998876. This almost perfect fit is indeed very poor if attention is paid to the pattern of residuals [+ + - - - - + + + + + + - -]. Systematic deviations can either indicate a systematic error in the experiment (which can not be tested since the details of the measurements are not known) or, as it turns out in this case, the use of an incorrect or inadequate model. The Claussius-Clapeyron equation does not exactly represent the vapor pressure data over a wide temperature range. Results similar to those

shown in Figure 6 are obtained by using a weighting factor transformation dependent (WLR) (Asuero and Bueno, 2011, Asuero and Gonzalez, 1989, Asuero and González, 2007, 1989, de Levie, 2012, 1986):

$$w_i = \frac{1}{\left(\frac{\partial Y_i}{\partial y_i}\right)} = \frac{1}{\left(\frac{\partial \ln P_i}{\partial P_i}\right)} = P_i^2 \tag{40}$$

Table 18. CO_2 vapor pressure data versus temperatura using Clausius-Clapeyron equation (*lnP = A + B/T*) (Nogle, 1993).

T	P	1/T	ln P	ln P est	Residual
216.550	5.110	0.005	1.631	1.637	-0.006
222.050	6.444	0.005	1.863	1.864	-0.001
227.606	8.043	0.004	2.085	2.083	0.002
233.161	9.921	0.004	2.295	2.291	0.003
238.717	12.099	0.004	2.493	2.490	0.003
244.272	14.623	0.004	2.683	2.679	0.003
249.828	17.508	0.004	2.863	2.860	0.002
255.383	20.788	0.004	3.034	3.034	0.001
260.939	24.510	0.004	3.199	3.199	0.000
266.494	28.702	0.004	3.357	3.358	-0.001
272.050	33.397	0.004	3.508	3.511	-0.002
277.606	38.636	0.004	3.654	3.657	-0.003
283.161	44.475	0.004	3.795	3.798	-0.003
288.717	50.939	0.003	3.931	3.933	-0.002
294.272	58.070	0.003	4.062	4.063	-0.001
299.828	65.916	0.003	4.188	4.188	0.000
304.161	72.768	0.003	4.287	4.283	0.005
	a1	-1988.945	10.822	a0	
	s(a1)	1.721	0.007	s(a0)	
	R2	1.000	0.003	s(y/x)	

The error lies not in the data, but in the model. We must try to improve the latter. A more general form of the equation is:

$$ln\ P = A + B/T + C\ ln\ T + D\ T \tag{41}$$

Results obtained in this later case (analysis by multiple linear regression) are much better (Table 19, Figure 7) than those obtained by the linear equation, with the residuals being distributed randomly. The standard deviation of the regression suggests that *ln P* can be calculated with an accuracy of 0.001, or an accuracy level of 0.1%. Another advantage is that T is used as a variable, rather than its inverse, so the interpolations become somewhat easier to calculate.

Table 19. CO_2 vapor pressure data versus temperatura using equation ($lnP = A + B/T + C \cdot lnT + D \cdot T$) (Nogle, 1993)

T	P	ln P	1/T	ln T	T	ln P est	y - yest
216.5500	5.1102	1.6312	0.0046	5.3778	216.5500	1.6311	0.000131641
222.0500	6.4439	1.8631	0.0045	5.4029	222.0500	1.8633	-0.000153221
227.6056	8.0430	2.0848	0.0044	5.4276	227.6056	2.0848	-2.3707E-05
233.1611	9.9211	2.2947	0.0043	5.4517	233.1611	2.2945	0.000162525
238.7167	12.0985	2.4931	0.0042	5.4753	238.7167	2.4934	-0.000306487
244.2722	14.6230	2.6826	0.0041	5.4983	244.2722	2.6825	0.000133521
249.8278	17.5082	2.8627	0.0040	5.5208	249.8278	2.8626	7.06843E-05
255.3833	20.7880	3.0344	0.0039	5.5428	255.3833	3.0346	-0.000193665
260.9389	24.5101	3.1991	0.0038	5.5643	260.9389	3.1991	-1.01222E-05
266.4944	28.7017	3.3570	0.0038	5.5854	266.4944	3.3568	0.000143651
272.0500	33.3968	3.5085	0.0037	5.6060	272.0500	3.5083	0.000149554
277.6056	38.6364	3.6542	0.0036	5.6262	277.6056	3.6541	7.43584E-05
283.1611	44.4747	3.7949	0.0035	5.6460	283.1611	3.7947	0.000204959
288.7167	50.9390	3.9306	0.0035	5.6654	288.7167	3.9305	8.55518E-05
294.2722	58.0702	4.0617	0.0034	5.6845	294.2722	4.0620	-0.00034836
299.8278	65.9159	4.1884	0.0033	5.7032	299.8278	4.1895	-0.001079349
304.1611	72.7681	4.2873	0.0033	5.7176	304.1611	4.2863	0.000965059
0.0354	-18.3653	-4353.3367	112.8273	a3	a2	a1	a0
0.0014	0.7395	94.8033	4.1040	s(a3)	s(a2)	s(a1)	s(a0)
1.0000	0.0004	#N/A	#N/A	R2	s(y/x)		
19138738.4636	13.0000	#N/A	#N/A	F	df		
11.2081	0.0000	#N/A	#N/A	SS(REG)	SS(RES)		

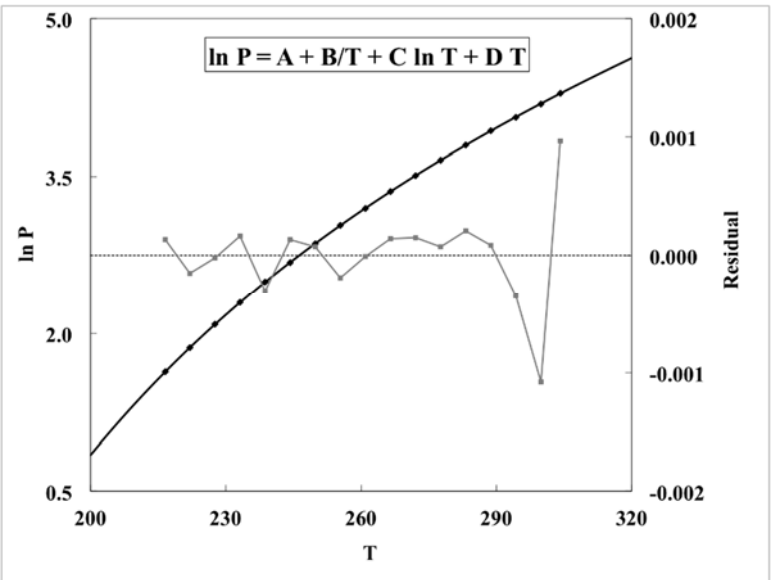

Figure 7. CO_2 vapor pressure as a function of temperature according to the expanded model $ln\ P = A + B/T + C\ ln\ T + D\ T$.

CONCLUSION

Linear regression is probably one of the most important problems in data analysis (Giloni et al., 2006). However, some problems appear in applying regression data analysis usually associated with mathematical statistical unfamiliarity aspects. Heterocedasticity (irregular, heterogeneous or non constant variance), when ignored, leads to an inadequate estimation and inference. In order to solve this problem, data transformation and weighted least squares regression can be used.

Varying quality data are easily handled with weighted least squares. Weights stem directly from the least squares criterion, i.e., the likelihood function, requiring variance known and independent of the model parameters, and weights equals to the reciprocal variance (Seber, 2003; Rawlings et al., 1998; Thompson, 1982; Draper and Smith, 1965; Williams, 1959). In real applications weights are almost never exactly known and estimates must be used instead (Seber, 2003; Williams, 1959;

Engineering Statistics Handbook), which supposes a common drawback. Theoretical models (Danzer and Currie, 1998) or statistical tests (Sayago and Asuero, 2004; Penninckx et al., 1996) help in taking the decision of weighting or not. We may assume that a smooth variance function account for the heterocedasticity (Tellinghuisen, 2005a; Davidian, 1990; Carroll and Ruppert, 1988), as an alternative procedure to replication.

Restore linearity (linearly transformable models) may be the aim of response transformation (Rawlings, 1998; Meloun and Militký, 1994) though sometimes pursue variance stabilization or getting normality. Heterogeneous variance may be removed (Natrella, 1963) as additional benefits besides linearity. In practice, however, it is difficult to find one simple transformation that simultaneously satisfies different criteria (Draper and Hunter 1969, Bartlett, 1947). A general scientific theory or some possible distributional assumptions or an empirical plotting of the data (Weisberg, 2005; AMC, 1994) often serve as the basis for these simple transformations (logarithmic, power, root). The adequacy of the transformed data model has to be checked before proceeding.

Box and Cox (1964) derived a general method for choosing transformations of the response valid both in simple and multiple linear regressions (Li and Moor, 2002; Lee et al., 1999). The Box-Cox original paper has been a fruitfully source of inspiration (Andersen et al., 1999; Mateu, 1997; Chinn, 1996; Sakia, 1992) that generated as much theoretical work as practical applications. Presence of outliers and heterocedasticity (Zarembka, 1974) adversely affect the robustness of Box-Cox method. However many advantages (Logothetis, 1990) are ascribed to this method: i) complete use of the information contained in the data at hand; ii) assurance of the validity of simple and normality assumptions; iii) shorter confidence interval for λ; and iv) treatment in absence of replication (Box and Meyer, 1986).

Keep the model as simple as possible (Rawlings et al., 1998; Bates and Watts et al., 1988; Garfinkel and Fegley, 1984) is a golden key in science because facilitates the understanding of the system as well as the communication of results. FDA guidelines for bioanalytical method validation follow this direction (Singtoroj et al., 2006) when states “the

selection of weighting and use of a complex regression equation should be justified."

REFERENCES

Aarons, L., Toon, S., Rowland, M., (1987). Validation of assay methodology used in pharmacokinetic studies. *J. Pharmacol. Methods* 17, 337-346.

Acton, F.S., (1959). *Analysis of Straight Line Data.* New York, USA: Wiley.

Almeida, A.M., Castel-Branco, M.M., Falcao, A.C., (2003). Linear regression for calibration lines revisited: weighting schemes for bioanalytical methods. *J. Chromatogr. B* 774, 215-222.

AMC, (1994). Is my calibration linear? *Analyst* 119(11), 2363-2366.

Anderson K.P., Snow, R.L., (1967). A relative deviation, least squares method of data treatment. *J. Chem. Educ.* 44, 756-757.

Asar, Ö., İlk, Ö., Dağ, O., (2017). Estimating Box-Cox power transformation parameter via goodness of fit tests. *Commun. Stat. Simul. Comput.* 46(1), 91-105.

Asnin, L.D., (2016). Peak measurement and calibration in chromatographic analysis. *Trends Anal. Chem.* 81, 51-62.

Asuero, A.G., González, G., de Pablos, F., Gomez Ariza, J.L., (1988). Determination of the optimum working range in spectrophotometric procedures. *Talanta* 35, 531-537.

Asuero A.G., González, A.G., (1989). Some observations of fitting a straight line to data. *Microchem. J.* 40, 216-225.

Asuero, A.G., Sayago, A., Gonzalez, A., (2006). The correlation coefficient: an overview. *Crit. Rev. Anal. Chem.* 36, 1-19.

Asuero, A.G., González G., (2007). Fitting Straight Lines with Replicated Observations by Linear Regression. III. Weighting Data. *Crit. Rev. Anal. Chem.* 37, 143-172.

Asuero, A.G., (2009). A Hiperbolic Sine Procedure for the Spectrophotometric Evaluation of Acidity Constants for Two-Step Overlapping Equilibria. *J. Anal. Chem.* 64, 1026-1030.

Asuero, A.G., Martín Bueno, J., (2011). Fitting Straight Lines with Replicated Observations by Linear Regression. IV. Transforming Data. *Crit. Rev. Anal. Chem.* 41, 36-69.

Badertscher, M., Pretsch, E., (2006). Bad results from good data. *Trends Anal. Chem.* 25, 1131-1138.

Bagur, M.G., Morales, S., López-Chicano, M., (2009). Evaluation of the environmental contamination at an abandoned mining site using multivariate statistical techniques—The Rodalquilar (Southern Spain) mining district. *Talanta* 80, 377-384.

Barnet, V., (2004). *Environmental Statistical Methods and Applications.* New York, USA: Wiley; 161-173.

Bartlett, M.S., (1947). The use of transformations. *Biometr.* 3(1), 39-52.

Barton, J.S., (2011). A Comprehensive Enzyme Kinetic Exercise for Biochemistry. *J. Chem. Educ.* 88, 1336-133.

Bates, D.M., Watts, D.G., (2007). *Nonlinear Regression Analysis and its Applications.* New York, USA: Wiley.

Baumann, K., Wätzig, H., (1995). Appropriate calibration functions for capillary electrophoresis. II. Heterocedasticity and its consequences. *J. Chromatogr. A* 700, 9-20.

Baumann, K., (1997). Regression and calibration for analytical separation techniques. Part II. Validation, weighted and robust regression. *Process Contr. Quality,* 10, 75-112.

Bayne C.K., Rubin, I.B., (1986). *Practical Experimental Designs and Optimization Methods for Chemists.* Deerfield Beach, Fl, USA: VCR Publishers, pp. 61-62.

Beaumont, S., Martin, J., Asuero, A.G., (2016). A Potentiometric Evaluation of Stability Constants of Two-Step Overlapping Equilibria via a Bilogarithmic Hyperbolic Cosine Method. *J. Anal. Sci. Methods Instrum.* 6, 33-43.

Belloto J.R.J., Sokolovski, T.D., (1985). Residual analysis in regression. *Am. J. Pharm. Educ.* 49, 295-303.

Boccio, M., Asuero, A.G., Sayago, A., (2007). Spectrophotometric Evaluation of Stability Constants of 1:1 Weak Complexes from Mole Ratio Data Using the Bilogarithmic Hyperbolic Cosine Method. *J. Anal. Chem.* 62, 840-844.

Boef, A.G.C., le Cessie, S., Dekkers, O.M., (2015). Models with Transformed Variables. Interpretation and Software. *Epidemiol.* 26, 16-17.

Bolster, C.H., Hornberge, G.M., (2007). On the Use of Linearized Langmuir Equations. *Nutrient Manag. Soil Plant Anal.* 71, 1796-1806.

Boumans, P.W.J.M., McKenna, R.J., Bosveld, M., (1981). Analysis of the limiting noise and identification of some factors that dictate the detection limits in a low-powder inductively coupled argon plasma system. *Spectrochim. Acta B* 36, 1031-1058.

Box, G.E.P., Cox, D.R., (1964). An analysis of transformations. *J. Royal Stat. Soc. Ser. B* 26 (2), 211-252.

Box, G.E.P., Draper, N.R., (1987). *Empirical-Model-Building and Response Surfaces*. New York, USA: Wiley.

Brasil, B., Bettencourt da Silva, R.J.N., Camõesb, M.F.G.F.C., Salgueiro, P.A.S., (2013). Weighted calibration with reduced number of signals by weighing factor modelling: Application to the identification of explosives by ion chromatography. *Anal. Chim. Acta* 804, 287-295.

Brownlee, K.A., (1984). *Statistical Theory and Methodology in Science and Engineering*. 2nd ed., Malabar, FL: Robert E. Krieger.

Brüggemann, L., Wennrich, R., (2011). Application of a special in-house validation procedure for environmental–analytical schemes including a comparison of functions for modelling the repeatability standard deviation. *Accred. Qual. Assur.* 16, 89-97.

Bubert, H., Klockenkämper, R., (1983). Precision-dependent calibration in instrumental analysis. *Fres. Zeitschrif Anal. Chem.* 316, 186-193.

Bysouth, S.R., Tyson, J.F., (1986). A comparison of curve fitting algorithms for flame absorption spectrometry. *J. Anal. At. Spectrosc.* 1(1), 85-87.

Cai, J., Liu, R., Sun, C., (2008). Logistic Regression Model for Isoconversional Kinetic Analysis of Cellulose Pyrolysis. *Energy Fuels* 22, 867-870.

Canavos, G.C., (1984). *Applied Probability and Statistical Methods.* Toronto, Canada: Little, Brown and Company.

Candioti, L.V., De Zan, M.M., Cámara, M.S., Goicoechea, H.C., (2014). Experimental design and multiple response optimization. Using the desirability function in analytical methods development. *Talanta* 124, 123-138.

Capitán-Vallvey, L.F., Arroyo-Guerrero, E., Fernández-Ramos, M.D., Cuadros-Rodríguez L., (2006). Logit linearization of analytical response curves in optical disposable sensors based on coextraction for monovalent anions. *Anal. Chim. Acta* 561, 156-163.

Carroll R.J., Ruppert, D., (1988). *Transformation and Weighting in Regresion.* London, England: Chapman & Hall. □

Chang, H.S., (1977). A computer program for Box-Cox transformation and estimation technique. *Econometrica* 45(7), 1741.

Chen, C., (2013). Evaluation of Equilibrium Sorption Isotherm Equations. *Open Chem. Eng. J.* 7, 24-44.

Chen, X., (2015). Modeling of Experimental Adsorption Isotherm Data. *Information* 6, 14-22.

Chinn, S., (1996). Choosing a transformation. *J. Appl. Stat.* 23(4), 395-404.

Chow S.-C., Liu, J.-P., (1995). In *Statistical Design and Analysis in Pharmaceutical Sciences.* New York, USA: Marcel Dekker.

Concheiro, M., Castaneto, M., Kronstrand, R., Huestis, M.A., (2015). Simultaneous determination of 40 novel psychoactive stimulants in urine by liquid chromatography-high resolution mass spectrometry and library matching. *J. Chromatogr. A* 1397, 32-42.

Connors, K.A., (1987). *Binding Constants, the Measurement of Molecular Complex Stability.* New York, USA: Wiley, 115.

Cornish-Bowden, A., (2014). Analysis and interpretation of enzyme kinetic data. *Perspect. Sci.* 1, 121-125.

Crabbe, M.J.C., (1982). An enzyme-kinetic program for desk-top computeres. *Comput. Biol. Med.* 12 (4), 263-283.

Dada, A.O., Olalekan, A.P., Olatunya, A.M., Dada, O., (2012). Langmuir, Freundlich, Temkin and Dubinin–Radushkevich Isotherms Studies of Equilibrium Sorption of Zn2+ Unto Phosphoric Acid Modified Rice Husk. *IOSR J. Appl. Chem.* 3, 38-45.

Daniel, C., Wood, F.S., (1980). *Fitting Equations to Data: Computer Analysis of Multifactor Data.* 2nd ed., New York, USA: Wiley.

Danzer, K., Currie, L.A., (1998). IUPAC Guidelines for calibration in analytical chemistry. Part 1. Fundamentals and single component calibration. *Pure Appl. Chem.* 70, 993-1014.

Davidian, M., Haaland, P.D., (1990). Regression and calibration with non constant error variance. *Chemometr. Intell. Lab. Systems,* 9, 231-248.

de Beer, J.O., Naert, C., Deconinck, E., (2012). The quality coefficient as performance assessment parameter of straight line calibration curves in relationship with the number of calibration points. *Accred. Qual. Assur.* 17 (3), 265-274.

de Brito J.A.A., Chettle, D.R., (2009). Calibration of 109Cd KXRF systems for in vivo bone lead measurements: weighted least-squares regression with different weighting functions. *Phys. Med. Biol.* 54, L45-L50.

de Brito, J.A.A., de Carvalho, M.L., Chettle, D.R., (2009). Calibration of 109Cd KXRF systems for in vivo bone lead measurements: the guiding role of the assumptions for least-squares regression in practical problem solving. *Phys. Med. Biol.* 54, 919-934.

De Galan L., van Dalen, H.P.J., Kornblum, G.R., (1985). Determination of strongly curved calibration graphs in flame atomic absorption spectrometry: comparison of manually drawn and computer calculated graphs. *Analyst* 110, 323-329.

de Levie, R., (1986). When, why, and how to use weighted least squares. *J. Chem. Educ.* 63, 10-15.

de Levie, R., (2000). Curve fitting least squares. *Crit. Rev. Anal. Chem.* 30, 59-74.

de Levie, R., (2001). *How to Use Excel in Analytical Chemistry and in General Scientific Data Analysis.* Cambridge, England: Cambridge University Press.

de Levie, R., (2004). Estimating precision in derived quantities. *J. Chem. Educ*. 9(2), 80-88.

de Levie, R., (2012). *Advanced Excel for Scientific Data Analysis*. 3th ed., Brunswick, Maine: Atlantic Academic.

Dekkers, A.L.M., Slob, W., (2012). Gaussian Quadrature is an efficient method for the back-transformation in estimating the usual intake distribution when assessing dietary exposure. *Food Chem. Toxicol.* 50, 3853-3861.

Deming, W.E., (1943). *Statistical Adjustment of Data.* New York, USA: Dover.

Denderz, N., Lehotay, J., (2012). Application of the van't Hoff dependences in the characterization of molecularly imprinted polymers for some phenolic acids. *J. Chromatogr. A* 1268, 44-52.

Desimoni, E., (1999). A program for the weighted linear least squares regression of unbalanced response arrays. *Analyst* 124, 1191-1196.

Desimoni, E., Brunetti, B., (2009). About estimating the limit of detection of heteroscedastic analytical systems. *Anal. Chim. Acta* 655, 30-37.

Dosne, A-G., Bergstrand, M., Karlsson, M.O., (2016). A strategy for residual error modeling incorporating scedasticity of variance and distribution shape. *J. Pharmacokinet. Pharmacodyn.* 43, 137-151.

Dowd, J.E., Riggs, D.S., (1965). A comparison of estimates of Michaelis-Menten kinetic constants from various linear transformations. *J. Biol. Chem.* 240 (2), 863-869.

Draper, N.R., Smith, H., (1998). *Applied Regression Analysis.* 3rd ed., New York, USA: Wiley.

du Toit, S.H.C., Steyn, A.G.W., Stumf, R.H., (1986). *Graphical Exploratory Data Analysis.* New York, USA: Springer Verlag.

El-Khaiary, M.I., (2008). xLeast-squares regression of adsorption equilibrium data: comparing the options. *J. Haz. Mat.* 158, 73-87.

El-Khaiary M.I., Malash G.F., Ho, Y-S., (2010). On the use of linearized pseudo-second-order kinetic equations for modeling adsorption systems. *Desalination* 257, 93-101.

Ellis K.J., Duggleby, R.G., (1978). What happens when data are fitted to the wrong equation? *Biochem. J.* 171, 513-517.

Elmore, D.T., Kingston, A.E., Shields, D.B., (1963). The computation of velocities and kinetic constants of reactions, with particular reference to enzyme-catalysed processes. *J. Chem. Soc.* 2070-2078.

Elskens, M., Baston, D.S., Stumpf, C., Haedrich, J., Keupers, I., Croes, K., Denison, M.S., Baeyens, W., Goeyens, L., (2011). CALUX measurements: Statistical inferences for the dose–response curve. *Talanta* 85, 1966-1973.

Engineering Statistics Handbook 4.1.4.3. Weighted least squares regression. http://www.itl.nist.gov/div.898/handbook/pmd/section1/pmd143.htm.

EPA, (2000). Guidance for Data Quality Assessment. *Practical Methods for Data Analysis*. EPA QA/G-9 QA00 Update, United States Environmental Protection Agency, EPA/600/R-96/084, 4-42.

EURACHEM/CITAC Guide, (2000). Quantifying Uncertainty in Analytical Chemistry, 2nd ed., http://www.measurementuncertainty.org/mu/guide/index.html.

FDA Home Page: "Guidance for Industry. Bioanalytical Method Validation." http://www.fda.gov/cder/guidance/ index.htm.

Foo, K.Y., Hameed, B.H., (2010). Insights into Modeling of Adsorption Isotherm Systems. *Chem. Eng. J.* 156, 2-10.

Gad, S.C., (1999). *Statistics and Experimental Design for Toxicologists*. 3th ed., Boca Raton, FL: CRC Press, 49-51.

Garden, J.S., Mitchell, D.G., Mills, W.N., (1980). Nonconstant variance regression techniques for calibration curve based analysis. *Anal. Chem.* 52, 305-307.

Garfinkel, D., Fegley, K.A., (1984). Fitting physiological models to data. *Am. J. Physiol. 246,* R641-R650.

Giloni, A., Simonof, J.S., Sengupta, B., (2006). Robust weighted LAD regression. *Comp. Stat. Data Anal.* 50, 3124-3140.

Goutelle, S., Maurin, M., Rougier, F., Barbaut, X., Bourguignon, L., Ducher, M., Maire, P., (2008). The Hill equation: a review of its capabilities in pharmacological modelling. *Fundam. Clin. Pharmacol.* 22, 633-648.

Gu, H., Liu, G., Wang, J., Aubry, A., Arnold, M.E., (2014). Selecting the correct weighting factors for linear and quadratic calibration curves with least-squares regression algorithm in bioanalytical LC-MS/MS assays and impacts of using incorrect weighting factors on curve stability, data quality, and assay performance. *Anal. Chem.* 86 (18), 8959-8966.

Hawker, D., (2015). Kinetics of Carbaryl Hydrolysis: An Undergraduate Environmental Chemistry Laboratory. *J. Chem. Educ.* 92, 1531-1535.

Heinzerling, P., Schrader, F., Schanze, S., (2012). Measurement of Enzyme Kinetics by Use of a Blood Glucometer: Hydrolysis of Sucrose and Lactose. *J. Chem. Educ.* 89, 1582-1586.

Herman, R.A., Scherer, P.N., Shan, G., (2008). Evaluation of logistic and polynomial models for fitting sandwich-ELISA calibration curves. *J. Immunolog. Methods* 339, 245-258.

Heydorn, K., Anglow, T., (2002). Calibration uncertainty. *Accred. Qual. Assur.* 7, 153-158.

Hladky, P.W., (2011). Chemical Dosing and First-Order Kinetics. *J. Chem. Educ.* 88, 776-781.

Howarth R., Thompson, M., (1976). Duplicate analysis in geochemical practice. Part 2. Examination of the proposed method and examples of its use. *Analyst* 101, 699-709.

Hoyle, M.H., (1973). Transformations—An introduction and a bibliography. *Int. Stat. Rev.* 41(2), 203-223.

Huang, C.-L., Moon, L.C., Chang, H.S., (1978). A computer program using the Box-Cox transformation technique for the specification of functional form. *Am. Stat.* 32(4), 144.

Hughes, H., Hurley, P.W., (1987). Precision and accuracy of test methods and the concept of K-factor in chemical analysis. *Analyst* 112, 1445-1449.

Hwang, L.-J., (1994). Impact of variance function estimation in regression and calibration. *Methods Enzymol.* 240, 150- 170.

Ingle, Jr. J.D., Crouch, S.R., (1972). Evaluation of precision of quantitative absorption spectrometric measurements. *Anal. Chem.* 44, 1375-1386.

Ingle, Jr. J.D., (1974). Precision of atomic absorption spectrometric measurements. *Anal. Chem.* 46, 2161-2171.

ISO 5725, (1994). *Accuracy (Trueness and Precision) of Measurement Methods and Results. Part 2: Basic Methods for the Determination of Repeatibility and Reproducibility of a Standard Measurement Method* (ISO, Geneva).

ISO11843–2, (2000). Capacity of Detection. Part 2. Metrology in the Linear Calibration Case. ISO, Geneva.

Jain, R.B., (2010). Comparison of three weighting schemes in weighted regression analysis for use in a chemistry laboratory. *Clin. Chim. Acta* 411, 270-279.

Johnson, K.J., (1980). *Numerical Methods in Chemistry*. New York, USA: Marcal Dekker, 245.

Jukic, D., Sabo, K., Scitovski, R., (2007). A review of existence criteria for parameter estimation of the Michaelis–Menten regression model. *Ann. Univ. Ferrara* 53, 281-291.

Jurs, P., (1970). Weighted least squares curve fitting using functional transformations. *Anal. Chem.* 42, 747- 750.

Jurs, P.C., (1986). *Computer Software Applications in Chemistry.* New York, USA: Wiley, 37-38.

Kapteyn, J.C., (1903). *Skew Frequency Curves in Biology and Statistics.* Groningen, The Netherlands: P. Noordhoff.

Kapteyn, J.C., van Uwen, M.J., (1916). *Skew Frequency Curves in Biology and Statistics.* Groningen, The Netherlands: Hoitrema Brothers.

Katskov, D., Hlongwane, M., Heitmann, U., Florek, S., (2012). High-resolution continuum source electrothermal atomic absorption spectrometry: Linearization of the calibration curves within a broad concentration range. *Spectrochim. Acta Part B* 71-72, 14-23.

Kemp, G.J., (1985). The susceptibility of calibration methods to errors in the analytical signal. *Anal. Chim. Acta* 176, 229-237.

Kirkup, L., Mulholland, M., (2004). Comparison of linear and non- linear equation for univariate calibration. *J. Chromatogr. A* 1029, 1-11.

Kleijburg, M.R., Pijpers, F.W., (1985). Calibration graphs in atomic-absorption spectrophotometry. *Analyst* 110, 147-150.

Klockenkämper, R., Bubert, H., (1986). Improvement of precision in spectrochemical analysis by correlation of intensity measurements. *Fres. Zeitschrif Anal. Chem.* 323, 112-116.

Komsta, Ł., (2010). A new general equation for retention modeling from the organic modifier content of the mobile phase. *Acta Chromatographica* 22.

Komsta, Ł., (2013). Statistical Evaluation and Validation of Quantitative Methods of Drug Analysis. Chapter 11. In: *Thin Layer Chromatography in Drug Analysis*. Komsta, L., Waksmundzka-Hajnos, M., Sherma, J., CRC Press, pp. 187-192.

Korany, M.A., Maher, H.M., Galal, S.M., Ragab, A.A., (2013). Comparative study of some robust statistical methods: weighted, parametric, and nonparametric linear regression of HPLC convoluted peak responses using internal standard method in drug bioavailability studies. *Anal. Bioanal. Chem.* 405 (14), 4835-4848.

Krupcık, J., Mydlova, J., Majek, P., Simon, P., Armstrong, D.W., (2008). Methods for studying reaction kinetics in gas chromatography, exemplified by using the 1-chloro-2,2-dimethylaziridine interconversion reaction. *J. Chromatogr. A* 1186, 144-160.

Lavagnini, I., Favaro, G., Magno, F., (2004). Non-linear and nonconstant variance calibration curves in analysis of volatile organic compounds for testing of water by the purge-and-trap method coupled with gas chromatography/mass spectrometry. *Rapid Commun. Mass Spectrom.* 18, 1383-1391.

Lavagnini, I., Favaro, G., Magno, F., (2005). Non-linear and non-constant variance calibration curves in analysis of volatile organic compounds for testing of water by the purge-and-trap method coupled with gas chromatography/mass spectrometry. *Rapid Comm. Mass Spectrom.* 18, 1383-1391.

Lavagnini, I., Magno, F., (2007). A statistical overview of univariate calibration, inverse regression, and detection limits: Application to gas chromatography/mass spectrometry technique. *Mass Spectrom. Rev.* 26(1), 1-18.

Lavagnini, I., Urbani, A., Magno, F., (2011). Overall calibration procedure via a statistically based matrix-comprehensive approach in the stir bar sorptive extraction–thermal desorption–gas chromatography–mass spectrometry analysis of pesticide residues in fruit-based soft drinks. *Talanta* 83, 1754-1762.

Lee, J.C., Chen, D-T., Hung, H-N., Chen, J.J., (1999). Analysis of drug dissolution data. *Stat. Med.* 18(7), 799-814.

Lee, J.-C., Ramsey, M.H., (2001). Modeling measurement uncertainty as a function of concentration: an example from a contaminated land investigation. *Analyst* 126, 1784-1791.

Leslie, D.S., Kohn, R., Nott, D.J., (2007). A general approach to heteroscedastic linear regression. *Stat. Comp.* 17, 131-146.

Li, B.B., Moor, B., (2002). The general Box-Cox transformation in multiple regression analysis. *Commun. Stat. Simul. Comput.* 31(4), 673-687.

Li, C., Liu, J., Di, D., Jiang, S., (2008). Analysis of Three Flavonoids in Oxytropis kansuensis Bunge by RP-LC–DAD Coupled with Weighted Least-Squares Linear Regression. *Chromatographia* 68, 773-779.

Logothetis, N., (1990). Box-Cox transformations and the Taguchi methods. *Appl. Stat.* 39(1), 31-48.

Mager, P.P., (1991). *Design Statistics in Pharmacochemistry*. New York, USA: Wiley, pp. 20-44.

Malaeb, Z.A., (1997). A SAS code to correct for non-normality and nonconstant variance in regression and ANOVA models using the Box-Cox method of power transformation. *Environ. Monit. Assess.* 47(3), 255-273.

Mansilha, C., Melo, A., Rebelo, H., Ferreira, I.M.P.L.V.O., Pinho, O., Domingues, V., Pinho, C., Gameiro, P., (2010). Quantification of endocrine disruptors and pesticides in water by gas chromatography-tandem mass spectrometry. Method validation using weighted linear regression schemes. *J. Chromatogr. A* 1217 (43), 6681- 6691.

Marengo, E., Robotti, E., Bobba, M., Righetti, P.G., (2008). Evaluation of the Variables Characterized by Significant Discriminating Power in the

Application of SIMCA Classification Method to Proteomic Studies. *J. Proteom. Res.* 7, 2789-2796.

Markovic, D.D., Lekic, B.M., Rajakovic-Ognjanovic, V.N., Onjia, A.E., Rajakovic, L.V., (2014). A New Approach in Regression Analysis for Modeling Adsorption Isotherms. *Sci. World J.* 1-17.

Mateu, J., (1997). Methods of assessing and achieving normality applied to environmental data. *Environ. Manag.* 21(5), 766-777.

McLean, A.M., Ruggirello, D.A., Banfield, C., Gonzalez, M.A., Bialer, M., (1990). Application of a variance-stabilizing transformation approach to linear regression of calibration lines. *J. Pharm. Sci.* 79(11), 1005-1008.

Meites, L., (1979). Some new techniques for the analysis and interpretation of chemical data. *Crit. Rev. Anal. Chem.* 8, 1-53.

Meloun, M., Militký, J., Forina, M., (1992). *Chemometrics for Analytical Chemistry. Vol. 1: PC-Aided Statistical Data Analysis.* New York, USA: Ellis Horwood, pp. 71-77.

Meloun, M., Pluharová, M., (2000). Thermodynamic dissociation constants of codeine, ethylmorphine and homatropine by regression analysis of potentiometric titration data. *Anal. Chim. Acta* 416, 55-68.

Meloun, M., Hill, M., Militký, J., Kupka, K., (2003). Assessment of the mean value of 17-hydroxypregnenolone in the umbilical blood of new-borns by the exploratory analysis of biochemical data. *Comp. Methods Programs Biomed.* 70(3), 187-197.

Meloun, M., Sanka, M., Nemec, P., Krıtkova, S., Kupka, K., (2005). The analysis of soil cores polluted with certain metals using the BoxeCox transformation. *Environ. Pollut.* 137, 273-280.

Meloun, M., Militký, J., (2011). Statistical Data Analysis: A Practical Guide. 1st Edition, Woodhead Publishing India, pp. 57-63.

Mermet, J-M., (2010). Calibration in atomic spectrometry: A tutorial review dealing with quality criteria, weighting procedures and possible curvatures. *Spectrochim. Acta Part B* 65, 509-523.

Miller-Ihli, N.J., O'Haver, T.C., Harnly, J.M., (1984). Calibration and curve fitting for extended range AAS. *Spectrochim. Acta* 39B (2-3), 1603-1614.

Miller, J.N., Miller, J.C., (2010). *Statistics and Chemometrics for Analytical Chemistry*. 6th ed., Harlow, England: Prentice-Hall.

Modamio, P., Lastra, C.F., Mariño, E.L., (1996). Determination of analytical error function for β-blockers as a possible weighting method for the estimation of the regression parameters. *J. Pharm. Biomed. Anal.* 14, 401- 408.

Mosteler, R., Tukey, J.W., (1977). *Data Analysis and Regression: A second course in statistics.* Reading, MA: Addison-Wesley.

Mullins, E., (2003). *Statistics for the Quality Control Laboratory.* Cambridge, England: RSC.

Muteki, K., Blackwood, D.O., Maranzano, B., Zhou, Y., Liu, Y.A., Leeman, K.R., Reid, G.L., (2013). Mixture Component Prediction Using Iterative Optimization Technology (Calibration-Free/Minimum Approach). *Ind. Eng. Chem. Res.* 52, 12258-12268.

Nascimento, R.S., Froes, R.E.S., Silva, N.O.C., Naveira, R.L.P., Mendes, D.B.C., Neto, W.B., Silva, J.B.B., (2010). Comparison between ordinary least squares regression and weighted regression in the calibration of metals present in human milk determined by ICP-OES. *Talanta* 80 (3), 1102-1109.

Natrella, M.G., (1963). The use of transformations, *Experimental Statistics*, National Bureau of Standards Handbook 91. Washington, DC: NBS, Chapter 20, pp. 201-203.

Naya, S., Cao, R., de Ullibarri I.L., Artiaga, R., Barbadillo, F., García, A., (2006). Logistic mixture model versus Arrhenius for kinetic study of material degradation by dynamic thermogravimetric analysis. *J. Chemometr.* 20(3-4), 158-163.

Noblitt, S.D., Berg, K.E., Cate, D.M., Henry, C.S., Characterizing nonconstant instrumental variance in emerging miniaturized analytical techniques. *Anal. Chim. Acta* 915, 64-73.

Noggle, N., (1993). *Practical Curve Fitting and Data Analysis: Software and Self-Instructions for Scientists and Engineers.* Chichester, England: Horwood.

O'Connell, M.A., Belanger, B.A., Haaland, P.D., (1993). Calibration and assay development using the four-parameter logistic model. *Chemometr. Intell. Lab. Systems* 20, 97-114.

Olivieri, A.C., (2015). Practical guidelines for reporting results in single- and multi-component analytical calibration: A tutorial. *Anal. Chim. Acta* 868, 10-22.

Oppenheimer, L., Capizzi, T.P., Weppelman, R.M., Mehta, H., (1983). Determining the lowest limit of reliable assay measurement. *Anal. Chem.* 55, 638-643.

Osborne, J.W., (2010). Improving your data transformations: applying the Box-Cox transformation. *Pract. Assess. Res. Eval. 15, 1-9.*

Osmari, T.A., Gallon, R., Schwaab, M., Barbosa-Coutinho, E., Baptista Severo Jr. J., Pinto, J.C., (2013). Statistical Analysis of Linear and Non-linear Regression for the Estimation of Adsorption Isotherm Parameters. *Adsorpt. Sci. Technol.* 31, 433-458.

Pardue, H.L., Hewitt, T.E., Milano, J.N., (1974). Photometric errors in kinetics and equilibrium analysis based on absorption spectroscopy. *Clin. Chem.* 20, 1028-1042.

Peace, K.E., (1988). *Biopharmaceutical Statistics for Drug Development.* New York, USA: Marcel Dekker, pp. 357-359.

Penninckx, W., Hartmann, D., Massart, D.L., Smeyers-Verbeke, J., (1996). Validation of the calibration procedure in atomic ab- sorption spectrometric methods. *J. Anal. At. Spectrom.* 11, 237-246.

Pereira da Silva, C., Soares Emídio, E., Rodrigues de Marchi, M.R., (2015). Method validation using weighted linear regression models for quantification of UV filters in water samples. *Talanta* 131, 221-227.

Phillips, L.J., Alexander, J., Hill, H.M., (1990). Quantitative Characterization of Analytical Methods, in Analysis for Drugs and Metabolites including Anti-infective Agents, E. & Reid I. D. Wilson, eds., London, England: RSC, pp. 23-36.

Piergiovanni, P.R., (2014). Adsorption Kinetics and Isotherms: A Safe, Simple, and Inexpensive Experiment for Three Levels of Students. *J. Chem. Educ.* 91, 560-565.

Prudnikov, E.D., Shapkina, Y.S., (1984). Random errors in analytical methods. *Analyst* 109, 305-307.

Rajakovic, L.V., Markovic, D.D., Rajakovic-Ognjanovic, V.N., Antanasijevic, D.Z., (2012). Review: The approaches for estimation of limit of detection for ICP-MS trace analysis of arsenic. *Talanta* 102, 79-87.

Rawlings, J.O., Pantula, S.G., Dickey, D.A., (1998). *Applied Regression Analysis. A Research Tool.* 2nd ed., New York, USA: Springer-Verlag.

Reigh, J.G., Wangermann, G., Rhode, K., Falck, M., (1972). General strategy for parameter estimation from isosteric and allosteric kinetic data and binding measurements. *Eur. J. Biochem.* 26, 368-379.

Rios, S., (1977). *Métodos Estadísticos*. 2nd ed., Madrid, España: Ediciones del Castillo.

Ritz, C., Baty, F., Streibig, J.C., Gerhard, D., (2015). Dose-Response Analysis Using R. *Plos One,* 1-13.

Rocke, M., Lorenzato, S., (1995). A two-component model for measurement error in analytical chemistry. *Technometr.* 37, 176-184.

Rocke, D.M., Durbin, B., Wilson, M., Kahn, H.D., (2003). Modeling uncertainty in the measurement of low-level analytes in environmental analysis. *Ecotoxicol. Environ. Saf.* 56, 78-92.

Rodbard, D., Frazier, G.R., (1975). Statistical Analysis of radioligand assay. *Methods Enzymol.* 37, 3-22.

Rodbard, D., Lenox, R.H., Wray, H.L., Ramseth, D., (1976). Statistical characterization of random errors in the radioimmunoassay dose-response variable. *Clin. Chem.* 22, 350-358.

Rode, R.A., Chinchilli, V.M., (1988). The use of Box-Cox transformations in the development of multivariate tolerance regions with applications to clinical chemistry. *Am. Stat.* 42(1), 23-30.

Rothman, L.D., Crouch, S.R., Ingle, Jr, J. D., (1975). Theoretical □and experimental investigation of factors affecting precision in molecular absorption spectrophotometry. *Anal. Chem.* 47, 1226-1233.

Roy, C., Chakrabarty, J., (2014). Quality by Design-Based Development of a Stability-Indicating RP-HPLC Method for the Simultaneous Determination of Methylparaben, Propylparaben, Diethylamino

Hydroxybenzoyl Hexyl Benzoate, and Octinoxate in Topical Pharmaceutical Formulation. *Sci. Pharm.* 82, 519-539.

Rudnyi, E.B., (1996). Statistical model of systematic errors: linear error model. *Chemometr. Intell. Lab. Systems* 34, 41-54.

Sadray, S., Rezaee, S., Rezakhah, S., (2003). Non-linear heterocedastic regression model for determination of methotrexate in human plasma by high performance liquid chromatography. *J. Chromatogr. B* 787, 293-302.

Safari, G.H., Nasseri, S., Mahvi, A.H., Yaghmaeian, K., Nabizadeh, R., Alimohammadi, M., (2015). Optimization of sonochemical degradation of tetracycline in aqueous solution using sono-activated persulfate process. *J. Environ. Health Sci. Eng.* 13, 1-15.

Sakia, R.M., (1992). The Box-Cox transformation technique—A review. *Statistician* 41(2), 169-178.

Sands, D.E., (1974). Weighting factors in least squares. *J. Chem. Educ.* 51, 473-474.

Santoyo, E., Guevara, M., Verma S.P., (2006). Determination of lanthanides in international geochemical reference materials by reversed-phase high-performance liquid chromatography using error propagation theory to estimate total analysis uncertainties. *J. Chromatogr. A* 1118, 73-81.

Sayago, A., Asuero, A.G., (2004). Fitting straight lines with replicated observations by linear regression: Part II. Testing for homogeneity of variances. *Crit. Rev. Anal. Chem.* 34, 133-146.

Sayago, A., Boccio M., Asuero, A.G., (2004). Fitting straight lines with replicated observations by linear regression: the least squares postulates. *Crit. Rev. Anal. Chem.* 34, 39-50.

Sayago, A., Asuero, A.G., (2006). Spectrophotometric evaluation of stability constants of 1:1 weak complexes from continuous variation data. *Int. J. Pharm.* 321, 94-100.

Schlesselman, J., (1971). Power families: A note on the Box and Cox transformation. *J. Royal Stat. Soc. B* 33(2), 307-311.

Schwartz, L.M., Gelb, R.I., (1978). Statistical analysis of titration data. *Anal. Chem.* 50, 1571-1576.

Sclove, S.L., (1972). (Y vs x) or (Log y vs x). *Technometr.* 14, 391.

Seber, G.A.F., (2003). Linear Regression Analysis. New York, USA: Wiley.

Seethapathy, S., Górecki, T., (2010). Polydimethylsiloxane-based permeation passive air sampler. Part II: Effect of temperature and humidity on the calibration constants. *J. Chromatogr. A* 1217, 7907-7913.

Shumway, R.H., Azari, R.S., Kayhanian, M., (2002). Statistical approaches to estimating mean water quality concentrations with detection limits. *Environ. Sci. Tecnol*. 36(15), 3345-3353.

Singh, K.P., Rai, P., Singh, A.K., Verma, P., Gupta, S., (2014). Occurrence of pharmaceuticals in urban wastewater of north Indian cities and risk assessment. *Environ. Monit. Assess.* 186, 6663-6682.

Singtoroj, T., Tarning, J., Annerberg, A., Ashton, M., Berqvist, Y., White, N. J., Lindegardh, N., Day, N.P.J., (2006). A new approach to evaluate regression models during validation of bioanalytical assays. *J. Pharm. Biomed. Anal.* 41, 219-227.

Smith E.D., Mathews, D.M., (1967). Least squares regression lines: Calculations assuming a constant percent error. *J. Chem. Educ.* 44, 757-759.

Sousa, J.A., Reynolds, A.M., Ribeiro, A.S., (2012). A comparison in the evaluation of measurement uncertainty in analytical chemistry testing between the use of quality control data and a regression analysis. *Accred. Qual. Assur.* 17, 207-214.

Steliopoulos, P., Stickel, E., Haas, H., Kranz, S., (2006). Method validation approach on the basis of a quadratic regression model. *Anal. Chim. Acta* 572, 121-124.

Sun, X.-Y., Singh, H., Millier, B., Warren, C.H., Aye, W.A., (1994). Noise, filters and detection limits. *J. Chromatogr. A* 687, 259-281.

't Lam, R.U.E., (2010). Scrutiny of variance results for outliers: Cochran's test optimized. *Anal. Chim. Acta* 659, 68-84.

Tan, A., Awaiye, K., Trabelsi, F., (2014). Impact of calibrator concentrations and their distribution on accuracy of quadratic

regression for liquid chromatography-mass spectrometry bioanalysis. *Anal. Chim. Acta* 815, 33-41.

Tellinghuisen, J., (2001). Statistical error propagation. *J. Phys. Chem. A* 105(15), 3917-3921.

Tellinghuisen, J., (2005a). Statistical error in isothermal titration calorimetry: variance function estimation from generalized least squares. *Anal. Biochem.* 343, 106-115.

Tellinghuisen, J., (2005b). Understanding Least Squares through Monte Carlo Calculations. *J. Chem. Educ.* 82, 157-166.

Tellinghuisen, J., (2007). Weighted least-squares in calibration: What difference does it make? *Analyst* 132, 536-543.

Tellinghuisen, J., (2008a). Least squares with non-normal data: estimating experimental variance functions. *Analyst* 133, 161-166.

Tellinghuisen, J., (2008b). Weighted least squares in calibration: The problem with using "quality coefficients" to select weighting formulas. *J. Chromatogr. B* 872, 162–166.

Tellinghuisen, J., (2009a). Least squares in calibration: weights, nonlinearity, and other nuisances. Chapter 10. *Methods Enzymol.* 454, 259-285.

Tellinghuisen, J., (2009b). The least-squares analysis of data from binding and enzyme kinetics studies: weights, bias, and confidence intervals in usual and unusual situations. Chapter 10. *Methods Enzymol.* 467, 599-529.

Tellinghuisen, J., (2009c). Weighting Formulas for the Least-Squares Analysis of Binding Phenomena Data. *J. Phys. Chem. B* 113, 6151-6157.

Tellinghuisen, J., (2009d). Variance function estimation by replicate analysis and generalized least squares: A Monte Carlo comparison. *Chemometr. Intell. Lab. Systems* 99, 138-149.

Tellinghuisen, J., (2010a). Least-squares analysis of data with uncertainty in x and y: A Monte Carlo methods comparison. *Chemometr. Intell. Lab. Systems* 103, 160-169.

Tellinghuisen, J., (2010b). Least-Squares Analysis of Phosphorus Soil Sorption Data with Weighting from Variance Function Estimation: A

Statistical Case for the Freundlich Isotherm. *Environ. Sci. Technol.* 44, 5029-5034.

Tellinghuisen, J., Bolster, C.H., (2011). Using R2 to compare least square fit models: when it must fail. *Chemometr. Intell. Lab. Systems* 105 (2), 220-222.

Tellinghuisen, J., (2015). Using Least Squares for Error Propagation. *J. Chem. Educ.* 92, 864-870.

Teunissen, P.J.G., Amiri-Simkooei, A.R., (2008). Least-squares variance component estimation. *J. Geodesy,* 82, 65-82.

Thompson, M., Howarth, R.J., (1973). Rapid estimation and control of precision by duplicate determinations. *Analyst* 98, 153-160.

Thompson, M., (1976). Duplicate analysis in geochemical practice. Part 1. Theoretical approach and estimation of analytical reproducibility. *Analyst* 101, 690-698.

Thompson, M., (1978). Dupan 3, a subroutine for the interpretation of analytical data in geochemical analysis. *Comp. Geosci.* 4, 333-340.

Thompson, M., Howarth, R., (1978). New approach to estimation of analytical precision. *J. Geochem. Explor.* 9, 23-30.

Thompson, M., (1982). Regression methods in the comparison of accuracy. *Analyst* 107, 1169-1180.

Thompson, M., (1988). Variation of precision with concentration in an analytical system. *Analyst* 112, 1579-1587.

Thompson, M., (2007) Why are we weighting? Anal. Methods Committee. *AMC technical brief No 27.*

Tomassone, R., Lesquoy, E., Miller, C., (1983). *La Regression, nouveaux regards sur une anciene methode statistique.* Paris, France: Masson, 15, 38.

Tukey, J.W., (1977). *Exploratory Data Analysis.* Reading, MA: Addison-Wesley.

van Loco, J., Hanot, V., Huysmans, G., Elskens, M., Degroodt, J.M., Beemaert, H., (2003). Estimation of the minimum detectable value for the determination of PCBs in fatty food samples by GC-ECD: A curvilinear calibration case. *Anal. Chim. Acta* 483(1-2), 413-418.

Vanatta, L.E., Coleman, D.E., (2007). Calibration, uncertainty, and recovery in the chromatographic sciences. *J. Chromatogr. A* 1158, 47-60.

Vélez, J.I., Corre, J.C., Marmolejo-Ramos, F., (2015). A new approach to the Box–Cox transformation. *Frontiers Appl. Mathemat. Stat.* 1, 1-10.

Wang, D., Zhuo, J-Q., Zhao, M-P., (2010). A simple and rapid competitive enzyme-linked immunosorbent assay (cELISA) for high-throughput measurement of secretory immunoglobulin A (sIgA) in saliva. *Talanta* 82, 432-436.

Watters, R.L., Carroll, R.J., Spiegelman, C.H., (1988). Heterocedastic calibration using analyzed reference materials as calibration standards. *J. Res. Nati. Bur. Stand.* (U.S.) 93, 264-265.

Weisberg, S., (2005). *Applied Linear Regression*. 3rd ed, New York, USA: Wiley.

Williams, E.J., (1959). Regression Analysis. New York, USA: Wiley.

Wilson, M.D., Rocke, D.M., Durbin, B., Kahn, H.D., (2004). Detection limits and goodness-of-fits measures for the two component model of chemical analytical error. *Anal. Chim. Acta* 509, 197-208.

Winefordner, J.D., Svoboda, V., Cline, L.J., (1970). Sources of noise in atomic absorption measurements. *Crit. Rev. Anal. Chem*. 1, 233-239.

Xu, W., Que Hee, S.S., (2006). Gas chromatography–mass spectrometry analysis of di-n-octyl disulfide in a straight oil metalworking fluid: Application of differential permeation and Box–Cox transformation. *J. Chromatogr. A* 1101, 25-31.

Yamada, K.T., (1992). Standard deviation in weighted least-squares analysis. *J. Mol. Spectrosc.* 156, 512-516.

Yaroshenko, I., Kirsanov, D., Kartsova, L., Sidorova, A., Borisova, I., Legin, A., (2015). Determination of urine ionic composition with potentiometric multisensor system. *Talanta* 131, 556-561.

Zarembka, P., (1974). Transformation of variables in econometrics, in *Frontiers of Econometrics*. Zarembka P. New York, USA: Academic Press, pp. 81-104.

Zeng, Q.C., Zbang, E., Tellinghuisen, J., (2008). Univariate calibration by reversed regression of heteroscedastic data: a case study. *Analyst* 133 (12), 1649-1655.

Zeng, Q.C., Zhang, E., Dong, H., Tellinghuisen, J., (2008). Weighted least squares in calibration: Estimating data variance functions in high-performance liquid chromatography. *J. Chromatogr. A* 1206, 147-152.

Zhang, C-Y., Chai, X-S., (2015). A novel multiple headspace extraction gas chromatographic method for measuring the diffusion coefficient of methanol in water and in olive oil. *J. Chromatogr. A* 1385, 124-128.

Zitter, H., God, C., (1971). Ermittlung, Auswertung und Ursachen von Fehlern bei Betriebsanalysen. *Zeitschrif Anal. Chem.* 255, 1-9.

Zorn, M.E., Gibbons, R.D., Sonzogni, W.C., (1977). Weighted least squares approach to calculating limits of detection and quantification by modeling variability as a function of concentration. *Anal. Chem.* 69, 3069-3075.

Zorn, M.E., Gibbons, R.D., Sonzogni, W.C., (1977). Weighted least squares approach to calculating limits of detection and quantification by modeling variability as a function of concentration. *Anal. Chem.* 69, 3069-3075.

In: Linear Regression
Editor: Vera L. Beck
ISBN: 978-1-53611-992-3
© 2017 Nova Science Publishers, Inc.

Chapter 2

REGRESSION THROUGH THE ORIGIN

Julia Martín and Agustín G. Asuero*
Department of Analytical Chemistry, Faculty of Pharmacy,
The University of Seville, Seville, Spain

ABSTRACT

Regression through the origin, a very interesting topic, has usually received a scarce attention in the bibliography. This model is also known as the no-intercept model. It is applied because of subject matter theory or either when other physical and material considerations are necessary to taken into account. An intensive bibliographical search has been carried out with the purpose of gathering the literature on the subject, which is widely scattered. Some about one hundredth and thirty references have been compiled, comprising about twenty monographs and fifty scientific journals, from varying fields, e.g., analytical, biological, clinical, chemometrical, educational, environmental, pharmaceutical, physico-chemical, and statistical. We will dealt systematically with the homocedastic condition, i.e., variance of y's independent of x, errors of y's accumulative, the heterocedastic case, i.e., variance or standard deviation proportional to x values, respectively, and orthogonal

* Corresponding Author address: Agustín G. Asuero, Department of Analytical Chemistry, Faculty of Pharmacy, University of Seville, Seville, Spain.

regression (error in both axes). The chapter also covers topics such as prediction (using the regression line in reverse), leverage, goodness of fit, comparison between models with and without intercept, uncertainty, polynomial regression models without intercept, and an overview of robust regression through the origin.

Keywords: least squares method, regression, origin

INTRODUCTION

Regression and related fitting methods have found wide use (Finney, 1996, Deming, 1968; Howard, 2001) in the field of natural and social sciences. Though linear least squares regression is probably the most widely used modeling statistical method, linear regression through the origin, in spite of its importance, has not received a great attention. There are occasions when it appears appropriate for a regression line (Bissell, 1992; Brownlee, 1984; Freund et al., 2006; Myers, 1986; Noggle, 1993; Ryan, 2008) to pass through the origin, i.e., for the true relation line to be

$$\eta = \beta_1 \xi \tag{1}$$

As the true variables η and ξ are unobservable, the model to be fitted is

$$Y_i = \beta_1 x_i + \varepsilon_i \tag{2}$$

where (Y_i, x_i) is the i*th* pair of associated x_i and Y_i values, i is the number of data point, β_1 is the model parameter to be estimated, and ε_i is the error associated with the measurement Y_i. This model is also called the no-intercept model (Chatterjee et al., 2012; Afifi and Azen, 1972). When it is known in advance that the intercept term is zero, then one has to impose this on the model (Rousseeuw, 2001). Regression through the origin is

applied due to underlying theory or either when other circumstances, i.e., physical and material considerations (Eisenhauer, 2003) are necessary to taken into account. Many practical applications can be found where the model given by Eqn (2) is more appropriate than the one with intercept added

$$Y = \beta_0 + \beta_1 x + \varepsilon \tag{3}$$

An example may be the regression of dose against area under the curve (AUC) in pharmacokinetic studies (Bonate, 2011).

The error term in Eqn (2) is assumed to be normally distributed with mean zero and unknown variance

$$\sigma_i^2 = \frac{\sigma^2}{w_i} \tag{4}$$

where σ is a constant (which may be absorbed into the unknown σ_i) and the weighting factors w_i's are known for all i, being inversely proportional to the variances.

The aim of this contribution is to offer a primer on the regression through the origin to analytical chemists and other related researchers interested in this subject. A number of applications have been compiled in tabular form on this respect. Figure 1 shows the number of papers published per year. Some fifty journals are cited from the fields of analytical and physical chemistry, chemometrics, clinical chemistry, ecology, educational chemistry, environmental chemistry, industrial hygiene, pharmacy, biology, and statistics. The authors apologize for those papers may have overlooked or inadvertently omitted. The most cited journals are shown in Figure 2.

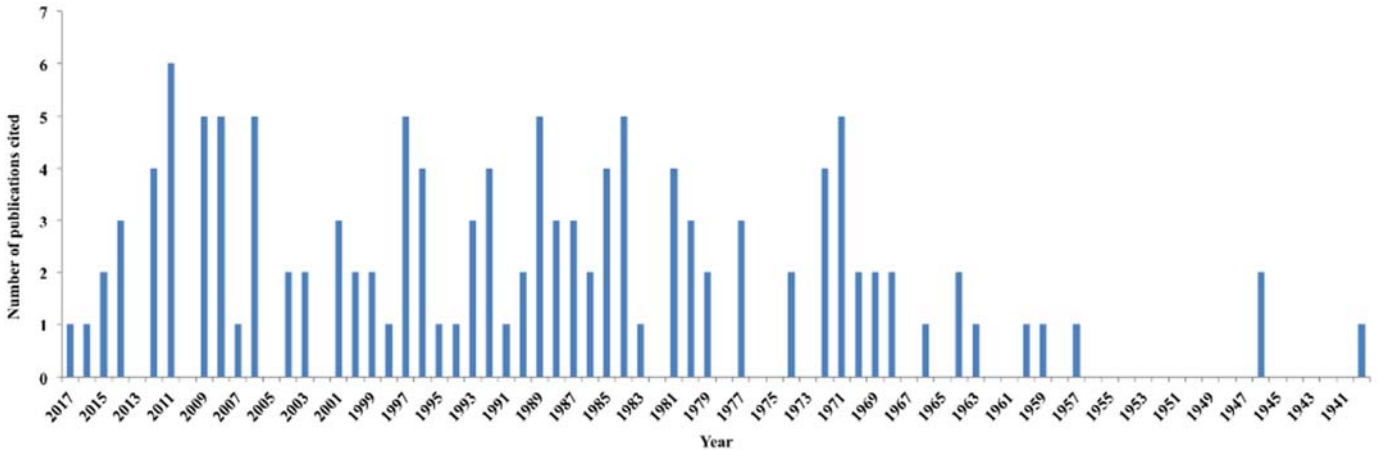

Figure 1. Number of publications cited per year.

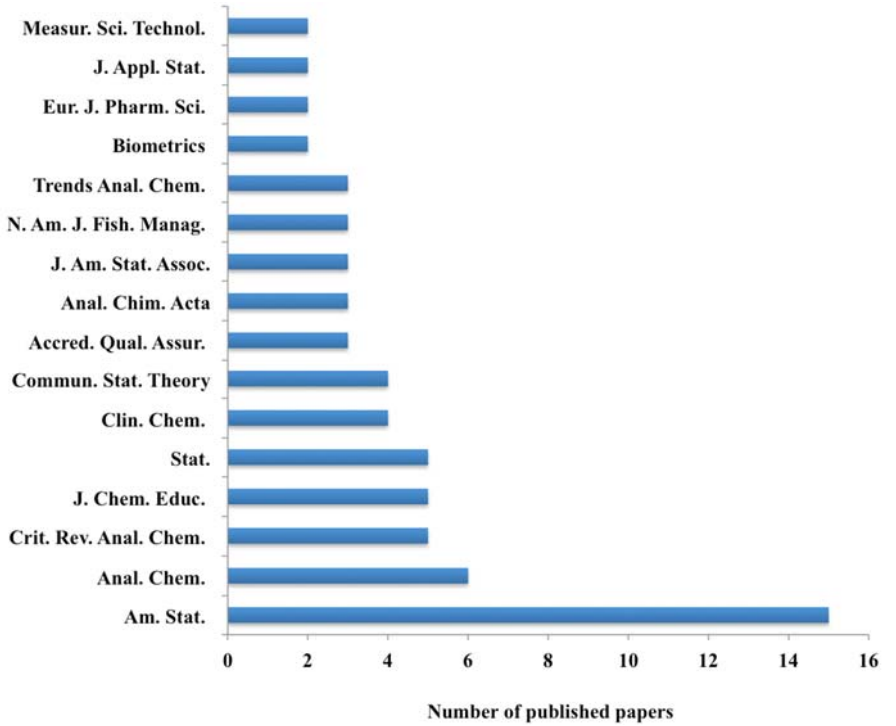

Figure 2. Number of papers cited by Journal.

WEIGHTED REGRESSION THROUGH THE ORIGIN

We proceed to derive the expression for weighted linear regression through the origin as in analytical chemistry is widely recognized (Asuero and Bueno, 2011; Asuero and Gonzalez, 2007; Sayago and Asuero, 2004) that the standard deviation increases with concentration, i.e., the variance is heterocedastic. The least squares method makes minimum the weighted sum of the residuals r_i

$$Q_{\min} = \sum w_i r_i^2 = \sum w_i \left(y_i - \hat{y}_i \right)^2 \tag{5}$$

of the *i*th fitted values

$$\hat{y}_i = b_1 x_i \tag{6}$$

where b_1 is the least squares estimates of β_1, obtained by making $\partial Q_{min}/\partial b_1=0$, the only normal equation from which

$$b_1 = \frac{\sum w_i x_i y_i}{\sum w_i x_i^2} \tag{7}$$

This formula is expressed in terms of the deviations from the origin unlike to the intercept case (Draper and Smith, 1977) that are expressed in terms of the deviation form the mean. Note that differentiating twice Q_{min} with respect to b_1 we obtain

$$\partial^2 Q / \partial b_1^2 = 2b_1 \sum w_i x_i^2 > 0 \tag{8}$$

which confirms Q_{min} to be a minimum. The sum of residuals is not necessarily equal to zero as occur for a model with intercept.

Several situations may be usually envisaged (Brownlee, 1984; Cox, 1971; Natrella, 1963; Turner, 1960) concerning weighting, as we will see in that follow.

Variance of y's independent of x

In those cases we get

$$w_i = 1 \tag{9}$$

and then

$$b_1 = \frac{\sum x_i y_i}{\sum x_i^2} \tag{10}$$

Variance proportional to x

Then we have (e.g., Georgian, 2009)

$$w_i = \frac{1}{x_i} \tag{11}$$

and

$$b_1 = \frac{\sum y_i}{\sum x_i} = \frac{\overline{y}}{\overline{x}} \tag{12}$$

Standard deviation proportional to x

The weight will be given by

$$w_i = \frac{1}{x_i^2} \tag{13}$$

which led to the slope value

$$b_1 = \frac{\sum \frac{y_i}{x_i}}{n} \tag{14}$$

Aston (1959) and Barlow (1989) follow a notation slightly different fro the one shown here.

STANDARD ERRORS IN REGRESSION THROUGH THE ORIGIN

By assuming that the values of x are free from error, applying the random error propagation law (Asuero et al., 1988) we get for the estimated variance of the slope

$$s_{b_1}^2 = \frac{s_{y/x}^2}{\sum w_i x_i^2} \tag{15}$$

The larger the term $\Sigma w_i x_i^2$, the greater the precision of the slope is. Note in addition that large values of $|x|$ will contribute substantially to this sum; increasing numbers of such values will also increase the denominator sum in Eqn. (15)

For the estimate of variance provided by the weighted residual we get

$$s_{y/x}^2 = \frac{\sum w_i \left(y_i - \hat{y}_i\right)^2}{n-1} = \frac{\sum w_i y_i^2 - b_1^2 \sum w_i x_i^2}{n-1} = \frac{\sum w_i y_i^2 - b_1 \sum w_i x_i y_i}{n-1} = \frac{\sum w_i y_i^2 - \frac{\left(\sum w_i x_i y_i\right)^2}{\sum w_i x_i^2}}{n-1} \tag{16}$$

In straight line regression through the origin $s_{y/x}^2$ has n-1 degrees of freedom (since only one parameter is estimated), not *n*-2 as is the case for a model with intercept. The rightmost expression in Eqn. (16) is the most convenient way (Green and Margerison, 1977) for computation purposes.

CONFIDENCE INTERVALS IN REGRESSION THROUGH THE ORIGIN

We get for the confidence interval for β_1

$$b_1 - t_{n-1,\alpha/2} s_{b_1} < \beta_1 < b_1 + t_{n-1,\alpha/2} s_{b_1} \tag{17}$$

From Eqn. (6) we have for the variance of a point on the true regression line

$$s_{\hat{y}_i}^2 = x_i^2 s_{b_1}^2 = \frac{x_i^2 s_{y/x}^2}{\sum w_i x_i^2} \tag{18}$$

Note that $s_{\hat{y}}$ increases linearly with x whereas $s_{\hat{y}} / \hat{y}$ (variation coefficient), remains constant.

The confidence interval for a point on the true regression line is then calculated (Bennett and Franklin, 1954) from

$$b_1 x_0 - t_{n-1,\alpha/2} \frac{x_0 s_{y/x}}{\sqrt{\sum w_i x_i^2}} \le x_0 \beta_1 \le b_1 x_0 + t_{n-1,\alpha/2} \frac{x_0 s_{y/x}}{\sqrt{\sum w_i x_i^2}} \tag{19}$$

The confidence band for the entire regression line is the region between two straight lines passing through the origin (Figure 3), whereas in a model with intercept they are parabolic curves. Thus, the interval becomes larger as we move away from the origin.

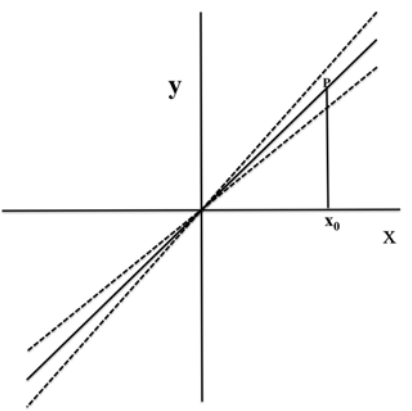

Figure 3. The solid line is the least squares line of slope b_1 passing through the origin and the set of points $(x_i, \hat{\mu}_i)$. The co-ordinates of the point P are, therefore, $(x_0, b_1 x_0)$. The dotted lines show the least squares estimates displaced vertically by ± one standard error, $s(\hat{\mu})$. The diagram shows clearly that the uncertainty associated with a least squares estimate of μ increases rapidly with increasing displacement of x_0 from the origin (Green and Margerison, 1977).

Hahn (1977) and Natrella (1963) deal with the confidence intervals to contain either b_1 or the true average response for a given x value, as well as the prediction interval containing a future response at a given x value. Hedayat (1970) and Hedayat et al. (1977) propose a test for detecting a monotonic relationship between the mean and variance. Iwase (1989) studies the case in which the y values follow an inverse Gaussian distribution being the coefficient of variation constant and unknown.

INVERSE EXTRAPOLATION: REVERSE USE OF THE REGRESSION LINE THROUGH THE ORIGIN

The mean $\overline{y}_0$ of m new observations allow to predict the corresponding x value and we may apply the Fieller's (1940) theorem to obtain the confidence limits for that prediction. Solving Eqn. (6) in reverse we get the point estimate

$$x_0 = \frac{\overline{y}_0}{b_1} \tag{20}$$

Here $\hat{x}_0$ is a non linear function of two normally distributed random variables: $\overline{y}_0$ and b_1, and will be approximately normally distributed about the true value of x, for which the m observations $\overline{y}_0$ were made, with (Bennett and Franklin, 1954; Lark et al., 1968), for the unweighted case (w_i=1)

$$s^2_{x_0} = \left(\frac{\partial x_0}{\partial \overline{y}_0}\right)^2 s^2_{y_0} + \left(\frac{\partial x_0}{\partial b_1}\right)^2 s^2_{b_1} = \frac{s^2_{y/x}}{b_1^2}\left(\frac{1}{m} + \frac{x_0^2}{\sum x_i^2}\right) \tag{21}$$

Although $\overline{y}_0 - x_0 b_1 = 0$, $y_0 - x_0 b_1 = z$, is not necessarily equals to zero since any particular $y_0$is not necessarily equal to the media of the m y_0 values. The difference, however, has a mean value of zero, and is normally distributed about zero. The ratio of z/s_z^2 is distributed as Student's t. The two capital premises had shown above support the Fieller's theorem (Bánfai, 2012; Bánfai and Kemény, 2012, Fieller, 1940, Schwartz and Gelb, 1984). Then we get

$$\frac{z^2}{s_z^2} = \frac{\left(\overline{y}_0 - b_1 x\right)^2}{s_z^2} = \frac{\left(\overline{y}_0 - b_1 x\right)^2}{s_{y/x}^2\left(\frac{1}{m} + \frac{x^2}{\sum x^2}\right)} = t_{n-1,\alpha/2}^2 \tag{22}$$

Confidence limits associated with x_0, at a prescribed level of significance α, valid even if the scatter of the y_0 values about the line is not small, can be calculated (Seber and Lee, 2003) by solving the quadratic equation

$$\left(b_1^2 - \frac{t^2 s_{y/x}^2}{\sum x^2}\right)x^2 - 2\overline{y}_0 b_1 x + \overline{y}_0^2 - \frac{t_{n-1,\alpha/2}^2 s_{y/x}^2}{m} = 0 \tag{23}$$

whose roots give the values of the lower and upper confidence limits of x, x_L and x_U, respectively. The difference between the upper and lower limits gives the confidence interval.

We may use a pooled variance (Cox, 1971; Seber and Lee, 2003) instead of $s_{y/x}^2$

$$s_p^2 = \frac{\sum_{i=1}^{n}\left(y_i - b_1 x_1\right)^2 + \sum_{j=1}^{m}\left(y_{oj} - \overline{y}_0\right)^2}{n + m - 2} \tag{24}$$

being in this case n+m-2 the degrees of freedom for the Student t.

GOODNESS OF FIT

We may be interested for example in fitting data as well as possible or in obtaining good predictions. Some difficulties appear in fitting non intercept models as usual statistics such as R^2 or F are not comparable with respect to the intercept model. A number of authors including Eisenhauer (2003), Gillingham and Heien (1971, Gordon (1981), Hahn (1977), Kozak and Kozak (1995), Hahn (1977) have addressed this issue that has generated a fruitful discussion around this subject, e.g., Beals (1972), Carmer and Walker (1971), D'Agostino (1971), Golsmith (1981), Gordon (1981b), Haws (1981), and Valentine (1971).

In the model with intercept, let

$$y_i - \overline{y} = \left(y_i - \hat{y}_i\right) + \left(\hat{y}_i - \overline{y}\right) \tag{25}$$

Summing and squaring over i we get

$$\sum\left(y_i - \overline{y}\right)^2 = \sum\left(y_i - \hat{y}_i\right)^2 + \sum\left(\hat{y}_i - \overline{y}\right)^2 + 2\sum\left(y_i - \hat{y}_i\right)\left(\hat{y}_i - \overline{y}\right) \tag{26}$$

and as the cross product in Eqn. (26) is equal to zero (Eisenhauer, 2003; Draper and Smith, 1997) we have

$$SST_m = SSE + SSR \tag{27}$$

The total sum of squares corrected for the mean (SST_m) may be partitioned into the residual error sum of squares (SSE) and the sum of squares due to regression (SSR). The coefficient of determination, R^2, is given by

$$R^2 = \frac{SSR}{SST_m} = 1 - \frac{SSE}{SST_m} \tag{28}$$

In the intercept term case R^2 is a measure of the total variability proportion that the regression explains. It coincides with the squared correlation coefficient ($-1 < R^2 < +1$) between x and y (or between y and y estimated) (Chatterjee et al., 2012; Draper and Smith, 1997).

When we deal with the regression through the origin, the cross product in Eqn. (26) will generally take a non zero value and therefore this equation is not valid as the basis for an analysis of variance. We may now write (Brownlee, 1984; Eisenhauer, 2003; Rousseeuw, 1987) for the non intercept model

$$y_i = \left(y_i - \hat{y}_i\right) + \hat{y}_i \tag{29}$$

and then

$$\sum y_i^2 = \sum \left(y_i - \hat{y}_i\right)^2 + \sum \hat{y}_i^2 + 2\sum \hat{y}_i \left(y_i - \hat{y}_i\right) \tag{30}$$

Taking into account that the cross product in Eqn. (30) is equal to zero, the (redefined) total sum of squares is decomposed now as

$$SST = SSE + SSR \tag{31}$$

(SSR being also redefined), and so

$$R^2 = 1 - \frac{SSE}{SST} \tag{32}$$

Equation (32) is the correct form of R^2 for regression through the origin (no intercept models). The R^2 values for models without an intercept are significantly higher than the ones for models without an intercept (Meloun and Militky, 2012; Meloun et al., 1994). Note, however, that the interpretations for the two formulas of R^2 (Eqns. (28) and (32)) are different. Fitting Eqn. (2) (non intercept model) by using the formula for R^2

given by Eqn. (28) may lead to a negative R^2 value, in some circumstances. For models without an intercept, no adjustment of Y is made.

A table of variance analysis corresponding to Eqn. (29) may be constructed as can be seen in Table 1. For details Brownlee (1984) should be consulted. Table 1 may be combined with the table of analysis of variance for regression with intercept giving a test of whether the model through the origin is or not adequate (Brownlee, 1984; Lark, 1968).

Table 1. Table of analysis of variance for regression through the origin (straight-line case)*

Source of variation	Sum of squares	Degrees of freedom	E[M.S.]**
Model (Due to line)	$\left(\sum w_i x_i y_i\right)^2 / \sum w_i x_i^2$	1	$\sigma^2 + \beta^2 \sum w_i x_i^2$
Residual	$\sum w_i \left(y_i - \hat{y}\right)^2$	$n-1$	σ^2
Total about origin	$\sum w_i y_i^2$	n	

* Adapted from Brownlee (1984); ** M.S. = mean squares.

COMPARISON BETWEEN MODELS WITH AND WITHOUT INTERCEPT

When dealing with straight lines relationships, the choice of choosing or not choosing between a model with and without an intercept can sometimes be posed. However, no simple solution to this problem (Casella, 1983; Gordon, 1981; Othman, 2014) is found. Applying linear regression with model intercept (Eqn. (3)) to a set of n data point (i=1,2,…n), we get for the estimated values of slope and intercept, b_1 and b_0, respectively

$$b_1 = \frac{S_{XY}}{S_{XX}} \tag{33}$$

$$b_0 = \overline{y}_w - b_1 \overline{x}_w \tag{34}$$

where S_{XX} and S_{YY} are the sum of the squares of deviations from the mean for the two variables x and y, respectively, and S_{XY} is the corresponding sum of the cross products (Asuero and Gonzalez, 2006; Asuero and Gonzalez, 1989; Martin and Asuero, 2017; Sayago and Asuero, 2004)

$$S_{XY} = \sum w_i x_i y_i - \frac{\left(\sum w_i x_i\right)\left(\sum w_i y_i\right)}{\sum w_i} \tag{35}$$

S_{XX} and S_{YY} may be easily derived from Eqn. (35) by substituting y_i by x_i, and x_i by y_i, respectively. The weighted (sample) mean values of x and y are given by

$$x_w = \frac{\sum w_i x_i}{\sum w_i} \tag{36}$$

and

$$y_w = \frac{\sum w_i y_i}{\sum w_i} \tag{37}$$

respectively. Note than when $w_i=1$ then $\Sigma w_i=n$.

In a model with intercept, the first normal equation (Eqn. (34)) requires that the fitted line pass thorough the centroid, and thus the weighted sum of residuals has to be zero. Nevertheless, in the case of non intercept term the sum of residual is not generally equals to zero.

Note that the slope of the regression through the origin given by Eqn. (7) may be expressed as

$$b_1 = \frac{S_{XY} + \frac{\left(\sum w_i x_i\right)\left(\sum w_i y_i\right)}{\sum w_i}}{S_{XX} + \frac{\left(\sum w_i x_i\right)^2}{\sum w_i}} = \frac{S_{XY} + \left(\sum w_i\right)\overline{x}_w \overline{y}_w}{S_{XX} + \left(\sum w_i\right)\overline{x}_w^2} \tag{38}$$

Thus, the line will not in general pass through the centroid so that Eqn. (7) and (38) are not equivalent.

In what follows from this section we use unweighted regression (w_i=1). Casella (1983) has shown that a new point (x_{n+1}, y_{n+1}) may be added to the previous full n set, forcing the straight line to pass now through the origin. Model based on Eqn. (2) is then applied, the slope being given by Eqn. (7). The new point added satisfies the identity

$$\left(x_{n+1}, y_{n+1}\right) = \left(n^* \overline{x}, n^* \overline{y}\right) \tag{39}$$

where

$$n^* = \frac{n}{\sqrt{n+1}-1} \tag{40}$$

The coordinates of the new point (its position with respect to the others) determine its leverage, that is, the amount of influence that has on each fitted value

$$h_{n+1} = \frac{1}{n+1}\left(1 + \frac{n^2 \overline{x}^2}{\sum_{i=1}^{n} x_i^2}\right) = \frac{1}{n+1}\left(1 + n\left(\frac{\overline{x}^2}{\frac{S_{XX}}{n} + \overline{x}^2}\right)\right) \tag{41}$$

Thus, the impact of the new (augmented) n+1 data point increases with

$\left(\overline{x}/\sqrt{(S_{XX}/n)}\right)^2$. The maximum discrepancy is expected to be when $\overline{x} >> \sqrt{(S_{XX}/n)}$. In those cases in which this new point is an outlier, the regression through the origin (Casellla, 1983; Meloun and Militky, 2011; Meloun et al., 1994) is not suitable.

On the other hand he corresponding standardized residual to the new (augmented) point is given by

$$r_{n+1}^{*} = \frac{b_0}{s_{y/x}\sqrt{1+\dfrac{\overline{x}^2}{S_{XX}}}} \tag{42}$$

where $s_{y/x}^2$ is the residual mean square from the full fit on the original n data points. Note that r_{n+1}^{*} is identical to the t statistic that tests H_0: $\alpha = 0$. It can be shown that

$$\left(r_{n+1}^{*}\right)^2 = \left(n-1\right)\left(\frac{s_{0,y/x}^2}{s_{y/x}^2}\right) - \left(n-2\right) \tag{43}$$

where $(s_{0,y/x})^2$ is the estimated residual mean squares from the regression through the origin and $(s_{y/x})^2$, as before, the estimated residual mean squares of the full regression. The $(r_{n+1}*)^2$ statistics is an exact measure of the relationship between the residual variances. The original paper of Casella (1983) should be consulted for additional details not included here.

For the leverage of models with intercept refer to Meloun et al. (1994) and Meloun and Miliktik (2012) for details. An excellent introduction to the topic is found in Sheater (2006).

Models through the origin may be used when consistency with the underlying theory or other adequate prior (material and physical) reasons are evident. There cases, however, where it is not clear which model should be used and the choice between both non-intercept and intercept models should be made with care. A comparison of the residual mean

squares obtained by the two models (Chatterjee et al., 2012) is a measure of the goodness of fit (closeness of observed and predicted values). Residual analysis may be also helpful in making decisions (Noggle, 1993). In addition if the intercept t test [$t = (b_0 - 0)/s_{b0}$] on the model given by Eqn. (3) is significant (reproducible non zero response at zero x value), this model should be used, otherwise use the regression through the origin given by Eqn. (2).

CAUTION CONCERNING ABOUT R^2

The coefficient of determination (R^2) is used to assess the goodness of fit for regression models. It is probably the single most extensively used measure of goodness of fit, but also widely misused (Asuero et al., 2006; Raposo, 2016; Scott and Wild, 1991), because the several alternative R^2 statistics are not generally equivalent (Kvalseth, 1985), except for linear models with an intercept term. Some regression packages compute R^2 in an inappropriate way (Hawkins, 1980; Gordon, 1981; Kozak and Kozak, 1995; Okunade et al., 1993). It is therefore possible to obtain different regression summary statistics, i.e., R^2 for the same equation specified in two equivalent ways (Uyar and Erdem, 1990). Becker and Kennedy (1992) possess a helpful exercise to understanding least squares remembering at the same time the problems with R^2 in those cases in which an intercept is not included in the regression model.

CONSTRAINED (CALIBRATION) EQUATIONS

In some instances, when dealing with fitting equations to data, the physical reality of the context dictates that the curve pass through one or more given points. Calibration curves, for example, often are known to pass through points such as (0, 0), (100, 100), (100, l), or (1, l) (Leary and Messick, 1985). As a matter of fact calibration curve may be forced to pass

through one or more independently selected points by using Lagrangian multipliers (Draper and Smith, 1997). Single ways of constraining a given equation (e.g., calibration) in order to pass through one or more independent selected points have been described by Meites and Leary (1985). The general case of forcing an equation to pass through a specific point or points is conveniently treated via the use of one or more Lagrange multipliers, which can supplement the more traditional least-squares curve fitting procedures. Some authors possess that constraints are appropriate because the fitting curve in each case is "known" to pass through the theoretical fixed points appealing to varying arguments. Helpful suggestions to decide that may whether or not to impose constraints have been advocated by Schwartz (1986) on the basis of the intended purpose of the regression. However, when the data are constrained to pass trough some fixed co-ordinates (x_0, y_0) point we may simply to shift the origin of coordinates system (Green and Margerison, 1977; Hahn, 1977) to the point (x_0, y_0) by means of the transformation

$$y' = y - y_0$$
$$x' = x - x_0$$

(44a,b)

STRAIGHT LINE IN THE CASE OF ACCUMULATIVE ERRORS

Sometimes, in engineering tests and in research, due to the characteristics of the experiment, the error related with each point involves the error related with all previous points. Then the successive observed values of the dependent variable y represent the cumulative magnitude (Mandel, 1964; Mandel, 1957; Natrella, 1963) of some successive effect at successive values of the independent variable x

$$y_i = \beta_1 x_i + \varepsilon_1 + \varepsilon_2 + ... + \varepsilon_{i-1} + \varepsilon_i \tag{45}$$

Then

$$y_i - y_{i-1} = \beta_1\left(x_i - x_{i-1}\right) + \varepsilon_i \tag{46}$$

which may be put in the form

$$z_i = \beta_1 L_i + \varepsilon_i \tag{47}$$

Note that ε_i's are statistically independent. By dividing through $\sqrt{L_i}$ we get

$$\frac{z_i}{\sqrt{L_i}} = \beta_1\sqrt{L_i} + \frac{\varepsilon_i}{\sqrt{L_i}} \tag{48}$$

equation which satisfies all requirement of ordinary least squares (errors statistically independent with zero mean and constant variance). Then the least squares (Eqn. 7) gives the estimate value of the slope

$$b_1 = \frac{\sum z_i}{\sum L_i} \tag{49}$$

from which the estimated value of its variance may be calculated (Mandel, 1964) as

$$s^2_{b_1} = \frac{1}{\left(\sum L_i\right)^2} k\sum L_i = \frac{\sum\left(\frac{d_i^2}{L_i}\right)}{\left(n-2\right)\sum L_i} \tag{50}$$

where

$$d_i = z_i - b_1 L_i \tag{51}$$

Mandel (1964) analyses Boyle's original data concerning P, V measurements and shows as the conventional treatment leads to non-randomness residual pattern, suggesting the model PV=constant (i.e., 1/P= V/constant) as incorrect. Nevertheless, a random residual pattern is obtained if cumulative errors are considered, showing the model of Boyle to be correct.

ERRORS IN VARIABLES METHOD

There are experiments in which the assumption of variable *x* free from error is not correct (Boccio et al., 2006; Duer et al., 2008; Sayago et al., 2004), being necessary in those cases to apply a more general treatment. We are going to consider three cases in order of increasing complexity.

Adcock's Regression though the Origin

Adcock, already in 1878, as indicated by Ripley and Thompson (1987) suggested minimizing the sum of the squares of the distance perpendicular to the line of adjustment. In the case of regression through the origin we get

$$r_i = \frac{y_i - b_1 x_i}{\sqrt{1 + b_1^2}} \tag{52}$$

and then

$$Q_{\min} = \frac{1}{1 + b_1^2} \sum \left(y_i - b x_i \right)^2 \tag{53}$$

By differentiating with respect to the slope and equation to zero

$$\frac{\partial Q_{\min}}{\partial b_1} = \frac{1}{1+b_1^2} \cdot 2\sum\left(y_i - b_1 x_i\right)\left(-x_i\right) + \left(\frac{-2}{\left(1+b_1^2\right)^2} \sum\left(y_i - b_1 x_i\right)^2 \right) \tag{54}$$

we get

$$b_1^2 + b_1\left(\frac{\sum x_i^2 - \sum y_i^2}{\sum x_i y_i} \right) - 1 = 0 \tag{55}$$

The minimization occurs when b_1 has the sign of $\Sigma x_i y_i$. This kind of regression is known with the name of orthogonal regression (equal errors in both axes).

Deming Regression

As the residuals are defined as

$$r_i = y_i - \hat{y}_i = y_i - b_1 x_i \tag{56}$$

their variance are given by

$$\sigma_{r_i}^2 = \sigma_{y_i}^2 + b_1^2 \sigma_{x_i}^2 - 2b_1 \operatorname{cov}\left(x_i, y_i\right) \tag{57}$$

and the we have the following general expression for the weights

$$w_i = \frac{1}{\sigma_{r_i}^2} = \frac{1}{\sigma_{y_i}^2 + b_1^2 \sigma_{x_i}^2 - 2b_1 \operatorname{cov}\left(x_i, y_i\right)} \tag{58}$$

If the (x_i, y_i) measurements are independent we have zero covariance and assuming the ratio of the variances of y_i to the x_i independent of the x values we obtain

$$w_i = \frac{1}{\sigma_{y_i}^2 + b_1^2 \sigma_{x_i}^2} = \frac{1}{\left(C + b_1^2\right)\sigma_{x_i}^2} \tag{59}$$

where

$$C = \frac{\sigma_{y_i}^2}{\sigma_{x_i}^2} \tag{60}$$

The weighted sum of the residuals has to be a minimum

$$Q_{\min} = \frac{1}{C + b_1^2} \sum \left(\frac{\left(y_i - b_1 x_i\right)^2}{\sigma_{x_i}^2} \right) \tag{61}$$

and then

$$\frac{\partial Q_{\min}}{\partial b_1} = \frac{1}{C + b_1^2} \left(2 \sum \left(\frac{y_i - b_1 x_i}{\sigma_{x_i}^2} \right) \left(-x_i\right) + \left(\frac{-2}{\left(C + b_1^2\right)^2} \right) \sum \left(\frac{\left(y_i - b_1 x_i\right)^2}{\sigma_{x_i}^2} \right) \right) = 0 \tag{62}$$

from which we get

$$b_1^2 + b_1 \frac{C\sum\left(\frac{x_i^2}{\sigma_{x_i}^2}\right) - \sum\left(\frac{y_i^2}{\sigma_{x_i}^2}\right)}{\sum\left(\frac{x_i y_i}{\sigma_{x_i}^2}\right)} - C = 0 \tag{63}$$

and like in the case of Eqn. (55) the minimization occurs when b_1 has the sign of the denominator of b_1 in Eqn. (63).

Weighted regression with weights given by Eqn. (59) but applied to models with intercept is known in the clinical literature (Linnet, 1993) with the name of Deming regression, and it is very used in comparison methods. It is also a kind of orthogonal regression, also named oblique regression.

Orthogonal Generalized Regression

We have assumed the independence of all the x_i and y_i, but when this is no the case we must to include cross terms involving the covariance of correlated variables. In addition, in those cases in which the ratio of variances of y to x values are not a constant

From Eqn. (5) in the most general case we get

$$\frac{\partial Q_{\min}}{\partial b_1} = \sum\left(w_i \frac{\partial r_i^2}{\partial b_1} + r_i^2 \frac{\partial w_i}{\partial b_1} \right) = 0 \tag{64}$$

and then

$$\sum w_i \frac{\partial r_i^2}{\partial b_1} = -\sum r_i^2 \frac{\partial w_i}{\partial b_1} \tag{65}$$

By differentiating r_i^2 and w_i with respect to b_1 we get, respectively

$$\frac{\partial r_i^2}{\partial b_1} = 2\left(y_i - b_1 x_i\right)\left(-x_i\right) = 2b_1 x_i^2 - 2x_i y_i \tag{66}$$

and

$$\frac{\partial w_i}{\partial b_1} = \frac{\partial}{\partial b_1}\left(\frac{1}{\sigma_{y_i}^2 + b_1^2\sigma_{x_i}^2 - 2b_1\operatorname{cov}\left(x_i, y_i\right)}\right) = \frac{-2a_1\sigma_{x_i}^2 + \operatorname{cov}\left(x_i, y_i\right)}{\left(\sigma_{y_i}^2 + b_1^2\sigma_{x_i}^2 - 2b_1\operatorname{cov}\left(x_i, y_i\right)\right)^2} = w_i^2\left(-2a_1\sigma_{x_i}^2 + \operatorname{cov}\left(x_i, y_i\right)\right) \tag{67}$$

By substituting Eqns. (67) and (66) into Eqn. (65) we get

$$\sum\left(w_i\left(2b_1 x_i^2 - 2x_i y_i\right)\right) = -\sum\left(r_i^2 w_i^2\left(-2b_1\sigma_{x_i}^2 + 2\operatorname{cov}\left(x_i, y_i\right)\right)\right) \tag{68}$$

which finally led to

$$b_1\sum w_i x_i^2 - \sum r_i^2 w_i^2\left(b_1\sigma_{x_i}^2 - \operatorname{cov}\left(x_i, y_i\right)\right) = \sum w_i x_i y_i \tag{69}$$

As the value of b_1 is depending on weighting factors and, at the same time, weighting factors are depending on b_1, it is necessary to use an iterative algorithm for solving the system. Thus the rigorous computation of the weights may become quite involved (Boccio et al., 2006; Sands, 1974). The starting value of b_1 Is obtained from Eqn. (7) (setting w_i=1). The new values of w_i are computed from Eqn. (58), and from these, an improved value of b1 is calculated applying Eqn. (69), an so on. A convergence criterion must be selected, e.g., k digits of b_1 should no be changed in the iteration

$$\left(\left(\frac{b_{1,n+1}}{b_{1,n}} \right) - 1 \right) < 10^{-k} \tag{70}$$

where n is the number of iterations. Once the optimum values of b1 and w_i's are known, sb12 is computed by applying the random error propagation law (Duer et al., 2008) to Eqn. (69).

The methodology followed to derive Eqn. (69) is that employed by Lisy et al. (1990), alternative procedure to the first derived by York (1969), for the most general case of intercept involved (covariance of the correlated variables included). This orthogonal generalized regression receives in the clinical bibliography the name (Martin, 2000) of general Deming regression. However, Martin (2000) follows an alternative derivation given by Williamson (1968).

ROBUST REGRESSION THROUGH THE ORIGIN

Departures from the assumptions inherent in single linear regression appear in the practice, exerting often a dramatic influence on the quality of statistical results. Robust regression leads to estimates that outliers do not influence so strongly as the standard least squares estimators. Thus, observations that lead to large residuals are down-weighted, i.e., weighted unequally. Robust regression methods are distribution free but require more computing than conventional least squares.

Least Median of Squares through the Origin (L_1 Regression)

The sum of the squared residuals may be replaced (Rousseeuw, 1984) by the median of squared residuals (Least Median of Squares, LMS, L_2 norm), making use of a (somewhat complex) non-linear optimization algorithm to carry out the necessary calculations, providing a robust version of the least squares regression. The PROGRESS (Program for

RObust reGRESSion) program (Rousseeuw and Leroy, 1987; Rousseeuw, 1988) has become popular on this respect, being available a most recent version (Rousseeuw and Hubert, 1997). Given the problems find in the PROGRESS software with the slope estimation when intercept lacks, Barreto and Maharry (2006) have devised an exact algorithm in the bivariate case, applicable in those circumstances. A new algorithm for a model with intercept suppressed including at most two unknown parameters covering bivariate cases have been set up by Kayhan and Gunay (2008), in the case of an odd number of data points. These later authors (Atilgan and Gunay, 2011) have also studied the LMS estimate for multiple linear regression models providing a more general algorithm. The problem may be treated as a convex optimization one. Note that robust methods are insensitive to departure from the normal distribution and to the outliers.

Least Absolute Deviation Regression through the Origin (L_1 Regression)

Minimizing the sum of absolute errors

$$\sum_{i=1}^{n} \left| \varepsilon_i \right| \tag{71}$$

is called L_1 regression (or L_1-norm regression (Draper and Smith, 1997).

A L_1 type estimator has been derived (Rieder, 1987) for regression through the origin (for both errors-in-variables and error-free-variables models). It is, among all estimators, minimax at finite sample size and extends Huber's (1964) robust interval estimator of location.

Deepest Regression through the Origin in Analytical Chemistry

The deepest regression method, a linear regression method, fits with the best depth relative to the data. Müller (2011) and Müller and Wellmann (2009) should be consulted for details to lengthy to include here. Deepest regression is reduced to

$$\hat{y}_i = median\left(\frac{y_i}{x_i}\right) \tag{72}$$

for a line through the origin, where observations with x_i=0 are not taken into account. Rousseeuw et al. (2001) showed a calibration data example for peak area in ng/ml for cadmium from graphite furnace atomic absorption spectrometry. The least squares line thorough the origin is displaced towards the outlier observed at the highest concentration standard, whereas the deepest regression through the origin is robust and fits the good data points.

POLYNOMIAL REGRESSION

No-intercept models more complex than the previous use seen here may also fitted to experimental data, e.g., a parabolic model passing through the origin (Hahn, 1977; Karl and Huber, 1997)

$$y = \beta_1 x + \beta_2 x^2 + \varepsilon \tag{73}$$

The law of Galileo on bodies in free fall (distance d travelled as a function of time t) is an example, $d=v_0\ t+(1/2)\ g\ t^2$, where the initial velocity v_0 is equal to β_1 and half of the acceleration of gravity is β_2

Solving by the least squares method we get (from the normal equations) the coefficients

$$b_2 = \frac{\left(\sum x_i y_i\right)\left(\sum x_i^3\right) - \left(\sum x_i^2 y_i\right)\left(\sum x_i^2\right)}{\left(\sum x_i^3\right)^2 - \left(\sum x_i^2\right)\left(\sum x_i^4\right)} \tag{74}$$

and

$$b_1 = \frac{\left(\sum x_i y_i\right) - b_2\left(\sum x_i^3\right)}{\left(\sum x_i^2\right)} \tag{75}$$

From signal values y_0 we may calculate the corresponding values of x, making a reverse use of the regression line

$$x_0 = -\frac{b_1}{2b_2} \pm \sqrt{\left(\frac{b_1}{2b_2}\right)^2 + \frac{y_0}{b_2}} \tag{76}$$

The sign of the root with physical meaning coincides with the sign of the parameter b_2, lacking of meaning the other root.

The variance of the regression will be given in this case by

$$s_{y/x}^2 = \frac{\sum\left(y_i - b_1 x_i - b_2 x^2\right)}{n-2} \tag{77}$$

Meites and Leary (1985) and Leary and Messick (1985) have treated constrained calibration curves with parabolic examples as shown above. Dalebrou (1974) reports variance analysis of polynomial regression with no intercept by means of the coefficients orthogonal method. D-optimal designs for polynomial regression models with no intercept have been the subject of statistical consideration (Fang, 2002).

ANALYTICAL APPLICATIONS OF REGRESSION THROUGH THE ORIGIN

Regression through the origin has found application in a variety of fields such as astronomy (Deming, 1968), computer tomography (Sun et al., 2000), ecology (Iwao, 1968; Waters et al., 2014), fishery (Bourgeois et al., 1997; Cade and Terrell, 1997; Bourgeois et al., 1996), forestry (Kozak and Kozak, 1995), industrial hygiene (Knight and Moore, 1987), parameter estimation (Cvetanovic et al., 1979) and wood science (Han, 1977). However, perhaps the largest applications have occurred within the framework of calibration (x free from error) in the field of analytical chemistry. Linear calibration functions passing through the origin have found use (Bánfai 2012a; Liteanu and Rica, 1980; Meloun and Militky, 2012; Mullins, 2003) in chromatographic, electrochemical, and other method of analysis. A number of authors dealing with that subject (no statistical journals) are included in Table 2.

For non-negligible x-errors the situation is more difficult to deal. The number of papers concerning specifically with regression through the origin with errors in the two variables is scarce. Andrews et al. (1996), Austin and Pelzer (1946), Kerrich (1966), Sands (1974), Tan and Jones (1989), Ripley and Thompson (1987), Synek (2001) and Winsor (1946) have been treated this topic. It has not been widely applied by chemists. Tan and Jones, for example, have been reported the relationship between the absorbance and chloride dioxide concentration (determined by iodometry). It could also been applied in the field of comparison methods, but in absence of systematic errors (Ripley and Thompson, 1987).

Analytical applications of regression through the origin are compiled in tabular form in Table 3.

Table 2. Authors who have published papers on calibration (regression linear through the origin with *x* free-from error) in non statistical journals

Alexander et al., 2015	Francis and Sobel, 1970	Kim and Burkart, 2008	Shayanfar and Shayanfar, 2011
Bonate, 2011	Georgian, 2009	Leroy and Messick, 1985	Strong III, 1979
Bánfai and Kemény, 2012	Hubert, 1997	Raposo et al., 2015	Synek et al., 2000
Bonate, 1992	Kemp, 1985	Ripley and Thompson, 1987	Van Zoonen et al., 1999
Dolan, 2009	Kemp, 1984a	Roy and Kas, 2014	
Ellerton and Strong, 1980	Kemp, 1984b	Schwartz, 1986	

Table 3. Some selected papers on regression through the origin

Content	Reference
The usage of R^2 as a measure of model fit and predictive power in QSAR or QSPR modelling. Suggestion of how to use it appropriately as a measure of model fit.	Alexander, Tropsha and Winkler, 2015
Methodology for developing priors from individual or combined meta-analyses which implicitly implies the assumption that there is variation around the meta-analytical relationships themselves. Examples of application to individual species are provided.	Hamel, 2015
Comparison between models with and without intercept and statement the beast one. Applying the method leverage point when a new point is added to the original data.	Abdulsalam Othman, 2014
The r_m^2 metrics and regression through origin approach: reliable and useful validation tools for predictive QSAR models (Commentary on 'Is regression through origin useful in external validation of QSAR models?').	Roy and Kar, 2014
Comparisson study of the proposed criteria using the regression through origin method (calculation with SPSS and Excel) for external validation and prediction capability for models developed using literature data. Prediction capability was evaluated using the statistically significant differences between absolute error values of training and test sets.	Shayanfar and Shayanfar, 2014

Table 3. (Continued)

Content	Reference
Iwao's patchiness regression through the origin: exploration whether fixing Iwao's m^*– m relation to go through the origin is theoretically justifiable, statistically advantageous given the methods used to estimate its parameters, and reduces the sample size required when used to design sequential sampling plans with no loss of sampling precision. Both analytical methods and resampling methods based on field data are employed.	Waters et al., 2014
Research on the suitability of interval hypotheses for a selection of analytical problems frequently occurring in the pharmaceutical setting. Overview of the statistical intervals and hypothesis tests used in the Dissertation. The interval hypothesis testing is discussed for the following topics: the transfer of analytical methods, the evaluation of the accuracy of analytical methods, the applicability of single-point calibration, and the content uniformity assessment.	Bánfai, 2012
Estimation of bias for single-point calibration using a proposed method based on the two one- sided tests (interval hypothesis). The test is performed by comparing a confidence interval for the bias to an allowable limit, defined in concentration units. Fieller's theorem was used for the ratio of two normally distributed random variables to construct the confidence interval for the bias.	Bánfai and Kemény, 2012
Survey of the development of different r_m^2 metrics followed by their applications in modeling studies for selection of the best QSAR models in different reports made by several workers.	Roy and Mitra, 2012
Clarification of the statement "*one often tends to use the origin point (0,0) in the data. However, whether that is best practice or not is entirely arguable.*" The argument that a zeroed instrument is expected to provide a point at (0,0) is specious and misleading.	Burkart and Kim, 2009
Calibration models: How to decide if a calibration curve goes through zero and some problems that can occur if the wrong choices are made.	Dolan, 2009
Evaluating 'goodness-of-fit' for linear instrument calibrations through the origin. A weighted regression coefficient is subsequently defined to evaluate the 'goodness-of-fit' and is expressed as function of the %RSD.	Georgian, 2009

Content	Reference
Properties of weighted least squares regression, particularly with regard to regression through the origin for establishment survey data, for use in periodic publications.	Knaub, 2009
The statistical reasons why regression through the origin should be used to analyze comparative data, and supports the recommendation of Garland et al. (1992) through additional geometric reasons.	Legendre and Desdevises, 2009
Discussing the visualization of statistical concepts and reply to the letter writed by Levie (2008) about including or not including an origin point (0,0) in a regression analysis for building a standard curve.	Kim and Burkart, 2008
The correct use of visualizing statistical concepts. Fails in the attempt of this test in the example described by Kim and Burkart, 2006 about "Beer's Law Plot."	Levie, 2008
Fitting curve passing for designated point to data for promoting the reproducibility of peripheral quantitative computed tomography.	Sun et al., 2008
A interactive and dynamic method of visual interactive regression minimizing the sum visible by allowing the individual to adjust heights in a bar graph. The interactive feature of Excel spreadsheet programs is utilized; use of the spinner bar is particularly helpful.	Kim and Burkart, 2006
Properties of the deepest regresion and applications in analytical chemistry: Regression through the origin, polynomial regression, the Michaelis–Menten model, and censored responses.	Rousseeuw et al., 2001
Linear regression of calibration lines passing through the origin was investigated for three models of y-direction random errors: normally distributed errors with an invariable standard deviation (SD) and log normally and normally distributed errors with an invariable relative standard deviation (RSD).	Synek, 2001
Uncertainties of mercury determinations in biological materials using an atomic absorption spectrometer. Study of potential sources of uncertainties as possible in order to work out a general model of determination of uncertainty in trace atomic absorption measurements.	Synek, Subrt and Marecek, 2000
Critical overview of most conflicting points concerning linear regression. Confidence bands and a discussion about the use of a line through the origin are included. In addition, the simplest expressions for expressing parameters to the appropriate significant figures from built-in calculator programs are also provided.	Giordano, 1999

Table 3. (Continued)

Content	Reference
Validation is put in the context of the process of producing chemical information. Two cases are presented in more detail: the development of a European standard for chlorophenols and its validation by a full scale collaborative trial, and the intralaboratory validation of a method for ethylene-thiourea using alternative analytical techniques.	Zoonen et al., 1999
Response to Comment of Cade and Terrell about cautions on Forcing Regression Equations through the Origin: This paper strengthens the caution to any-one considering no-intercept models for improving relations between fish density and weighted usable area. The explanations given by Cade and Terrell (1997) convincingly reinforce this warning.	Bourgeois et al., 1997
Comment to Cautions on Forcing Regression Equations through the Origin (Bourgeois et al., 1997): Prediction of biological response still depends largely of an detailed understanding of local biological conditions. Authors urged caution in forcing regression of fish density on weighted usable area through the origins, when such a forcing is contemplated, one should verify the calculations used by commercial statistical packages to generate summary statistics.	Cade and Terrell, 1997
Improved calibration for wide measuring ranges and low contents. For some calibrations, a straight line through the origin instead of a general straight line should be determined by regression analysis: advantages and restrictions.	Karl and Huber, 1997
A least-squares-based method for determining the ratio between two measured quantities.	Moreno, 1997
Relationship between principal components analysis and weighted linear regression for bivariate data sets: Application to linear, two-dimensional data sets with a zero intercept.	Andrews et al., 1996
The problem of fitting a straight line when both variables are subject to error. A brief review of the literature is undertaken, and one fitting method, the geometric mean functional realationship, is spotlighted and illustrated with two sets of example data.	Draper et al., 1991
Several methods of obtaining the best straught line from data in which the two variables are subject to errors of measurement are proposed and discussed.	Tan and Jones, 1989

Content	Reference
Statistical analysis techniques to compare pairs of dust samples: A straight line through the origin, linear with intercepts, logarithmic, a logarithmic (weigh + constant), and a fifth forced through the origin. It is suggested that the best estimate of the relation between two dust samplers can be obtained by a least squares determination of the straight line through the origin using transformed variables.	Knight and Moore, 1987
A regression-like technique, maximum-likelihood fitting of a functional relationship (MLFR), is explained and is demonstrated to work well. Under some conditions weighted regression provides a good approximation to MLFR, and so can be used if more convenient.	Ripley and Thompson, 1987
Suggestions that may be helpful to researchers in deciding whether or not to impose constraints.	Schwartz, 1986
The four main calibration methods (single separate or added standard and multiple separate or added standards) and some modifications are described mathematically and subjected to error-propagation analysis, to examine the likely effects of errors in the analytical signal on the overall accuracy and precision of the concentration estimate.	Kemp, 1985
Constrained Calibration Curves: How the use of Lagrange multipliers can supplement the more traditional least-squares curve fitting procedures. The concept of degrees of freedom when describing the variability of data around a calibration curve is also discussed.	Leary and Messick, 1985
Simple ways are described of constraining a calibration or other equation so that it will pass through one or more independently selected points and also give the “best” representation of any number of experimental data in terms of the model selected.	Meites and Leary, 1985
Theoretical aspects of one-point calibration: causes and effects of some potential errors, and their dependence on concentration.	Kemp, 1984a
New ways of using data from analytical-recovery studies to assess analytical nonlinearity, without access to samples of known concentration. A recovery-based method of assessing constant, proportional, and non-linear errors with use of as little as one sample pool of known concentration is described. In each case, the theoretical basis of the method and an outline of a practical experimental protocol is presented.	Kemp, 1984b

Table 3. (Continued)

Content	Reference
Evaluation of both qualitatively and quantitatively the bias error caused by an single-point-ratio calculations from an assumed linear response curve through zero for the case where the true response curve is a straight line with a significant intercept.	Cardone and Palermo, 1980
Comments on the correspondence about regression through the origin (Strong, 1979): Precision and accuracy shoud be considered from a statistical viewpoint and discussed.	Ellerton, 1980
Response of Strong to comments of Ellerton (1980) on regression through the origin: Strong agree thoroughly with Ellerton's definitions of precision and accuracy which apply to a chemist's repetitions of determinations on the same sample. Regarding the use of n-2 to calculate s_e, rather than n-1, as recommended by Ellerton, Strong felt that requiring the best straight line to pass through the origin was a constraint on the system and therefore constituted a reduction in the number of degrees of freedom.	Strong III, 1980
Determination of the precision and accuracy of kinetic data. Suggestions for the presentation of kinetic results and their uncertainties due to random and systematic errors. Regarding random errors, least-squares expressions are summarized, and confidence limits, propagation of errors, and change of variable are discussed. Sources of systematic errors are outlined, along with potential methods for their detection and estimation.	CvetanoviE, Singleton and Paraskevapoulos, 1979
Practical examples of fitting regression models with no intercept term. Caution in the use of the model is advised.	Hahn, 1979
Demostration of how in a photometric experiment if one measures the absorbances, *y*, of solutions having solute concentrations x, and if the solutions are expected to conform with Beer's law, one should fit a straight line that passes through the point $y = 0$ at $x = 0$. Strong proposes to accomplish this by using a single-parameter model equation, $y = b_1x$, rather than the conventional two-parameter model $y = a + b_2x$. The single-parameter model effectively forces the straight line to pass through (0, 0), but in either case, the slope, b_1 or b_2,represents the absorptivity. This will unavoidably reduce the precision slightly, but could increase the accuracy.	Strong III, 1979

Content	Reference
Weighting factors in least squares: When there is great variation among the variances, the assumption of constant weights can produce gross errors. A prcatical pH example.	Sands, 1974
Interval Estimate of the Ratio of an Unknown to a Standard. Methods for testing the suitability of the models under discussion are given.	Francis and Sobel, 1970

QSAR or QSPR: Quantitative Structure-Activity/Property Relationships (QSAR or QSPR)

FINAL COMMENTS

Some automated analytical systems, e.g., chromatographs, carrying out non-intercept fitting on routine basis (Mullins, 2003), mainly using single point calibration (Bánfai and Kemény, 2012), i.e., measuring only one standard and drawing the line from the origin to this measured point. Caution is required working on this way (Raposo, 2016) because chance component in the measure influences the line in a mayor way. It is essential in those cases in the validation step to investigate carefully the linearity of the response. The regression model thorough the origin, when applicable, gives estimates more précises (Hahn, 1977) than the ones obtained by the most usual model with the intercept.

It should be noted that special models should be adopted only for adequate prior reasons. Eisenhauer (2003) has said about regression through the origin that "it remains a subject of pedagogical neglect, controversy and confusion," concluding with regard to the practice of statistics that it "remains as much as art as it is science," being "the development of the statistical judgment as important as the computational ability.

A primer on regression through the origin has been reported in this contribution with the hope of being useful in teaching and in research.

REFERENCES

Acton, F.S., (1959). *Analysis of Straight Line Data: The equation y=b x.* New York, USA: Wiley, pp. 16-17.

Afifi, A.F., Azen, S.P., (1972). *Statistical Analysis, A Computer Oriented Approach.* 2nd ed., 1st ed., New York, USA: Academic Press, p.125; pp. 88-89.

Alexander, D.L., Tropha, A., Winkler, D.A., (2015). Beware of R^2: simple, unambiguous assessment of the prediction accuracy of QSAR and SSPR models. *J. Chem. Inf. Model.* 55(7), 1316-1322.

Andrews, D.T., Chen, L., Wentzell, P.D., Hamilton, D.C., (1996). Comments on the relationship between principal components analysis and weighted linear regression for bivariate data sets. *Chemometr. Intell. Lab. Systems* 34, 231-244.

Asuero, A.G., Bueno, J., (2011). Fitting straight lines with replicated observations by linear regression. IV. Transforming data. *Crit. Rev. Anal. Chem*. 41(1), 36-69.

Asuero, A.G., Gonzalez, G., (1989). Some observations on fitting a straight line to data. *Microchem. J*. 40(2), 216-225.

Asuero, A.G., Gonzalez, G., (2007). Fitting straight lines with replicated observations by linear regression. III. Weighting data. *Crit. Rev. Anal. Chem*. 37(3), 143-172.

Asuero, A.G., Gonzalez, G., de Pablos, F., Ariza, J.L.G., (1988). Determination of the optimum working range in spectrophotometric procedures. *Talanta* 35(7), 531-537.

Asuero, A.G., Sayago, A., Gonzalez, A.G., (2006). The correlation coefficient: an overview. *Crit. Rev. Anal. Chem*. 36(1), 41-59.

Atilgan, Y.K., Gunay, S., (2011). Least median of squares solution of multiple linear regression models through the origin. *Commun. Stat. Theory Methods* 40(22), 4125-4137.

Austen, A.E.W., Pelzer, H., (1946). Linear curves of best fit. *Nature* 157, 693-694.

Bánfai, B., (2012). *Statistical Problems in the Pharmaceutical Analysis*. PhD. Thesis, Budapest University of Technology and Economics, Budapest, pp. 43-56.

Bánfai, B., Kemény, S., (2012). Estimation of bias for single-point calibration. *J. Chemometrics* 26, 117-124.

Barlow, R., (1989). *Statistics. A Guide to the Use of Statistical Methods in the Physical Sciences*. New York, USA: Wiley, pp. 98-99.

Barreto, H., Maharry, D., (2006). Least median of squares and regression through the origin. *Comput. Stat. Data Anal.* 50(6), 1391-1397.

Beals, R.E., (1972). Regression through origin –comment. *Am. Stat.* 26(1), 54.

Becker, W., Kennedy, P., (1992). A lesson in least squares and R Squared. *Am. Stat.* 46(4), 282-283.

Bennett, C.A., Franklin, N.L., (1954). *Statistical Analysis in Chemistry and the Chemical Industry*. New York, USA: Wiley, pp. 232-234.

Bissell, A.F., (1992). Lines through the origin- is NO INT the answer?. *J. Appl. Stat.* 19(2), 193-210.

Boccio, M., Sayado, A., Asuero, A.G., (2006). A bilogarithmic method for the spectrophotometric evaluation of stability constants of 1:1 weak complexes from mole ratio data. *Int. J. Pharm*. 318, 70-77.

Bonate, P.L., (1992). Concepts in calibration theory. 2. Regression through the origin –when should it be used. *LC GC Magazine Sep. Sci.* 10(5), 378-379.

Bonate, P.L., (2011). Linear models and regression, In *Pharmacokinetic-Pharmacodynamics Modeling and Simulation*. Springer Science+Business Media, Chapter 2, pp. 61-100.

Bourgeois, G., Cunjak, R.A., Caissie, D., El-Jabi, N., (1996). A special and temporal evaluation of PHAB-SIM in relation to measured density of juvenile Atlantic salmon in a small stream. *N. Am. J. Fish. Manag.* 16, 154-166.

Bourgeois, G., Cunjak, R.A., Caissie, D., El-Jabi, N., (1997). Cautions on forcing regression equations through the origin: response to comments. *N. Am. J. Fish. Manag.* 17(1), 227-228.

Brownlee, K.A., (1984). Statistical Methodology in Science and Engineering. In *Regression through the origin*. 3rd ed., Malabat, Fla; R.G. Krieger, pp. 358-362.

Cade, B.S., Terrell, J.W., (1997). Comment: cautions on forcing regression equations through the origin. *N. Am. J. Fish. Manag.* 17 (1), 225-232.

Carmer, S.G., Walker, W.M., (1971). Regression through origin. *Am. Stat.* 25(5), 57-58.

Casella, G., (1983). Leverage and regression through the origin. *Am. Stat.* 37(2), 147-152.

Chatterjee, S., Hadi, A.S., Price, B., (2012). *Regression Analysis by Example: Regression through the origin*. 5th ed., New York, USA: Wiley, pp. 46-48.

Cox, C.P., (1971). Interval estimation for X-predictions from linear Y on X regression lines through the origin. *J. Am. Stat. Assoc.* 66(336), 749-751; (1972), 67(337), 252 (Erratum).

Cvetanovic, R.J., Singleton, O.L., Paraskcvopoulos, G., (1979). Evaluation of the mean value, and standard errors of rate constants and their temperature coefficients. *J. Phys. Chem.* 83(1), 50-60.

Dalebrou, M.A., (1974). Polynomial regression through the origin – analysis of variance by method of orthogonal coefficients. *Annales de l'amélioration des plantes* 24(1), 71-76.

David, H.A., (1972). Regression through origin –comment. *Am. Stat.* 26(1), 54.

de Levie, R., (2008). Visualizing statistical concepts. *J. Chem. Educ.* 85(5), 635.

Deming, T.J., (1968). The analysis of linear correlation in Astronomy. *Vistas Astron.* 10, 125-142.

Dolan, J.W., (2009). Calibration curves, Part 1: to b or not to b?. *LC GC Eur.* 190, 192-194.

D'Agostini, R.B., (1971). Regression through origin. *Am. Stat.* 25(5), 59.

Duer, W.C., Ogren, P., Meetze, A., Kitchen, C.J., von Lindern, R., Yaworsky, D.C., Boden, C., Gayer, J.A., (2008). Comparison of ordinary, weighted, and generalized least squares straight line

calibrations for LC-MS-MS, GC-MS, HPLC, GC, and enzymatic assays. *J. Anal. Toxicol*., 32(5), 329-338.

Draper, N.R., Smith, H., (1998). *Applied Regression Analysis*. 3rd ed., New York, USA: Wiley, pp. 121-233.

Eisenhauer, J.G., (2003). Regression thorough the origin. *Teach. Stat.* 25(3), 76-80.

Ellerton, R.W., Strong III, F.C., (1980). Comments on regression through the origin. *Anal. Chem.* 52(7), 1152-1154.

Fang, Z., (2002). D-optimal designs for polynomial regression models through the origin. *Stat. Probabil. Lett.* 57 (4), 343-351.

Fieller, E.C., (1940). The biological standardization of insulin. *J. Roy. Stat. Soc. Supp.* 7(1), 1-64.

Finney, D.J., (1996). A note on the history of regression. *J. Appl. Stat.* 23(5), 515-558.

Francis, M., Sobel, E., (1970). Interval estimate of the ratio of an unknown to a standard. *Anal. Chem.* 42(3), 314-320.

Freund, J.R., Wilson, W.J., Sa, P., (2006). *Regression Analysis, Statistical Modely of a Response Variable.* 2nd ed., Burlington, MA: Elsevier.

Georgian, T., (2009). Evaluating 'goodness-of fit' for linear instrument calibrations through the origin. *Int. J. Enviro. Anal. Chem.* 89, 383-388.

Gillingham, G., Heien, D., (1971). Regression through the origin. *Am. Stat.* 25(1), 54-55.

Giordano, J.L., (1999). On reporting uncertainties of the straight-line regression parameters. *Eur. J. Phys*. 20(5), 345-349.

Goldsmith, P.L., (1981). Letter to the Editor. *Stat*. 30(3), 234.

Gordon, H.A., (1981a). Errors in computer packages. Least squares regression through the origin. *Stat.* 30(1), 23-29.

Gordon, H.A., (1981b). Letter to Editor. *Stat*. 30(4), 305-308.

Green, J.R., Margerison, D., (1977). *Statistical Treatment of Experimental Data: The straight line through the origin or through some other fixed point.* Amsterdam: Elsevier, Chapter 12, pp. 198-235.

Hahn, G.J., (1977). Fitting regression models with no intercept term. *J. Qual. Technol.* 9(2), 56-61.

Hamel, O.S., (2015) A method for calculating a meta- analytical prior for the natural mortality rate using multiple life history correlates. *ICES J. Mar. Sci.* 72(1), 62-69.

Hawkins, D.M., (1980). A note on fitting a regression without an intercept term. *Am. Stat.* 34(4), 233.

Haws, A.P., Gordon, H.A., (1981). Letter to Editor. *Stat.* 30(4), 304-308.

Hedayat, A., (1970). Examination and analysis of residuals, diagnostic checking of residuals for detecting a special type of heteroscedasticity in linear regression through the origin. *Biometrics* 26(3), 603. (Joint Meeting of ENAR with IMS and ASA, Chape Hill, North Caroline).

Hedayat, A., Raktoe, B.L., Talwar, P.P., (1977). Examination and analysis of residuals: a test for detecting a monotonic relation between mean and variance in regression through the origin. *Commun. Stat. Theory Methods* 6(6), 497-506.

Howarth, R.J., (2001). A history of regression and related model-fitting in the earth sciences. *Nat. Resourc. Res.* 10(4), 241-286.

Huber, M.K.W., (1997). Improved calibration for wide measuring ranges and low contents. *Accred. Qual. Assur.* 2(8), 367-374.

Huber, P.J., (1964). Robust estimation of a location parameter. *Ann. Math. Stat*. 35, 73-101.

Iwao, S., (1968). A new regression method for analyzing the aggregation pattern of animal population. *Res. Popul. Ecol.* 10(1), 1-20.

Iwase, K., (1989). Linear regression through the origin with constant coefficient of variation for the inverse Gaussian distribution. *Commun. Stat. Theory Methods* 18(10), 3587-3593.

Kayhan, Y., Gunay, S.M., (2008). A new approach to least median of squares and regression through the origin. *Commun. Stat. Theory Methods* 37(5), 773-781.

Kemp, G.J., (1985). The susceptibility of calibration methods to errors in the analytical signal. *Anal. Chim. Acta* 176, 229-247.

Kemp, G.J., (1984). Theoretical aspects of one-point calibration: causes and effects of some potential errors, and their dependence on concentration. *Clin. Chem.* 30(7), 1163-1167.

Kemp, G.J., (1984). Assessment of analytical bias: four new ways to use recovery measurements. *Clin. Chem.* 30(7), 1168-1170.

Kerrich, J.E., (1966). Fitting the line y=ax when errors of observation are present in both variables. *Am. Stat.* 20(1), 24.

Kim, M-H., Burkart, M., (2008). The author replies. Including or not including an original point (0,0) in a regression analysis for building a standard curve. *J. Chem. Educ.* 85(5), 635-636.

Kim, M-H., Burkart, M., Kim, M.H., (2006). A method of visual interactive regression. *J. Chem. Educ.* 83(12), 1884.

Knaub, J.R., (2009) Properties of weighted least squares regression for cutoff sampling in establishment surveys. Conference paper Cuttof Sampling and Establishment Surveys. *InterStat J.* December.

Knight, G., Moore, E., (1987). Comparison of dust samplers: statistical analysis techniques. *Am. Ind. Hyg. Assoc. J.* 48(4), 344-353.

Kozak, A., Kozak, R.A., (1995). Notes on regression through the origin. *Forest. Chron.* 7(3), 326-330.

Kvalseth, T.O., (1985). Cautionary note about R^2. *Am. Stat.* 39(4), 279-285.

Lark, P.D., Craven, B.R., Bosworth, R.C.L., (1969). *The Handling of Chemical Data* (pp 159-163). Oxford, England: Pergamon Press.

Leary, J.J., Messick, E.B., (1985). Constrained calibration curves: a novel application of Lagrange multipliers in analytical chemistry. *Anal. Chem.* 57(4), 956-957.

Legendre, P., Desdevises, Y., (2009). Independent contrasts and regression through the origin. *J. Theoret. Biol.* 259(4), 727-743.

Linnet, K., (1993). Evaluation of regression procedures for method comparison studies. *Clin. Chem.* 39(3), 424-432.

Lisy, J.M., (1990). Multiple straight-line least squares analysis with uncertainties in all variables. *Comp. Chem*. 14, 189-192.

Liteanu, C., Rica, I., (1980). *Statistical Theory and Methodology of Trace Analysis*. New York, USA: Ellis Horwood, pp. 161-162.

Mandel, J., (1957). Fitting a straight line to certain type of cumulative data. *J. Am. Stat. Assoc.* 12(280), 552-566.

Mandel, J., (1964). *The Statistical Analysis of Experimental Data*. New York, USA: Dover, pp. 295-303.

Martin, J., Asuero, A.G., (2017). Weighting and transforming data in linear regression. In *Linear Regression: Models, Analysis and Applications*. Nova Science Publishers.

Martin, R.F., (2000). General Deming regression for estimating systematic bias and its confidence interval in method comparison studies. *Clin. Chem*. 46, 100-104.

Meites, L., Leary, J.J., (1985). Simple procedures for obtaining constrained calibration equations. *Anal. Chim. Acta* 176, 249-251.

Meloun, M., Militky, J., (2011). *Statistical Data Analysis. A Practical Guide with 1250 exercises and answer key on CD*. New Delhi: Woodhead Publ., pp 483-486.

Meloun, M., Militky, J., Forina, M., (1994). *Chemometrics for Analytical Chemistry. Volume 2: PC-Aided Regression and Related Methods*. New York, USA: Ellis Horwood, pp. 30-33.

Moreno, C., (1996). A least-squares based method for determining the ratio between two measured quantities. *Measur. Sci. Technol.* 7(2), 137-141; (1997), 8(8), 951.

Muller, C.H., (2011). Data depth for simple orthogonal regression with application to crack orientation. *Metrika* 74(2), 135-165.

Muller, C.H., Wellmann, R., (2009). *Data Depth for Classical and Orthogonal Regression.* Cors09 International Conference on Robust Statistics, Book of Abstracts (pp. 110-111), Università degli Studi di Parma, Facoltà di Economia, Riani, M., Ceroli, A., Agostinelli, C., Perrotta, D. Eds., Libero Libri- Claudio Agostinelli, Paese (TV), Itali.

Mullins, E., (2003). *Statistics for the Quality Control Chemistry Laboratory*. Cambridge, England: Royal Society of Chemistry (RSC), pp. 270-275.

Myers, R.H., (1986). *Classical and Modern Regression with Applications.* Boston, USA: Duxbury Press.

Natrella, M.G., (1963). *Experimental Statistics, NBS Handbook 91*. Washington: U.S. Government Printing Office, pp. 5-24 to 5-27.0, 483-485.

Noggle, J.H., (1993). *Practical Curve Fitting and Data Analysis. Software and Self-Instruction for Scientists and Engineers*. Chichester, England: Ellis Horwood.

Okunade, A.A., Chang, C.F., Evans, R.D., (1993). Comparative analysis of regression output summary statistics in common statistical packages. *Am. Stat.* 47(4), 298-303.

Othman, S.A., (2014). Comparison between models with and without intercept. *Gen. Math. Notes* 21(1), 118-127.

Raposo, F., (2016). Evaluation of analytical calibration based on least-squares linear regression for instrumental techniques: a tutorial review. *Trends Anal. Chem.* 77, 167-185.

Rieder, H., (1989). A finite-sample minimax regression estimator. *Stat.* 20(2), 211-221.

Ripley, B.D., Thompson, M., (1987). Regression techniques for the detection of analytical bias. *Analyst* 112(4), 377-383.

Rousseeuw, P.J., (1984). Least median of squares regression. *J. Am. Stat. Assoc.* 79(12), 871-880.

Rousseeuw, P., (1988). PROGRESS: a program for robust regression. *Trends Anal. Chem*. 7(9), 320-321.

Rousseeuw, P.J., Hubert, M., (1997). Recent development in PROGRESS. In *Lectur Notes-Monograph Series.* Vol. 31, Institute of Mathematical Statistics (IMS).

Rousseeuw, P.J., van Aelst, S., Rambali, B., Smeyers-Verbeke. J., (2001). Deepest regression in analytical chemistry. *Anal. Chim. Acta* 446, 245-256.

Rousseeuw, P.J., Leroy, A.M., (1987). *Robust Regression & Outlier Detection: Simple regression through the origin*. New York, USA: Wiley, pp. 62-65.

Roy, K., Kar, S., (2014). The r(m)(2) metrics and regression through origin approach: reliable and useful validation tools for predictive QASR models commentary on 'Is regression through the origin useful in external validation of QASR models? *Eur. J. Pharm. Sci.* 62, 111-114.

Roy, K., Mitra, I., (2012). On the use of the metric rm2 as an effective tool for validation of QASR models in computational drug design and predictive toxicology. *Mini Rev. Med. Chem.* 12(6), 491-504.

Ryan, T.P., (2008). *Modern Regression Methods.* 2nd ed., New York, USA: Wiley.

Sands, D.E., (1974). Weighting factors in least squares. *J. Chem. Educ.* 51(7), 473-474.

Sayago, A., Boccio, M., Asuero, A.G., (2004). Fitting straight lines with replicated observations by linear regression: the least squares postulates. *Crit. Rev. Anal. Chem*. 34(1), 39-50.

Sayago, A., Asuero, A.G., (2004). Fitting straight lines with replicated observations by linear regression. Part II. Testing for homogeneity of variances. *Crit. Rev. Anal. Chem*. 34(3-4), 133-146.

Schwartz, L.M., (1986). Effect of constraints on precision of calibration analyses. *Anal. Chem.* 58(1), 246-250.

Schwartz, L.M., Gelb, R.I., (1984). Statistical uncertainties of end points at intersecting straight lines. *Anal. Chem*. 56(8), 1487-1492.

Scott, A., Wild, C., (1991). Transformations and R^2. *Am. Stat.* 45(2), 127-129.

Seber, G.A.F., Lee, A.J., (2003). *Linear Regression Analysis.* 2nd ed., New York, USA: Wiley, p. 149.

Shayanfar, A., Shayanfar, S., (2011). Is regression through origin useful in external evaluation of QASR models?. *Eur. J. Pharm. Sci.* 87, 271-273.

Sheather, S.J., (2009). *A Modern Approach to Regression with R*. New York: Springer, pp 51-70, pp 115-123.

Strong III, F.C., (1979). Regression line that starts at the origin. *Anal. Chem.* 51(2), 298-299.

Sun, L., Xie, T., Fan Y.M., Zhang, C., (2008). *Fitting curve passing through designated point to data for promoting the reproducibility of peripheral quantitative computed tomography (pQCT).* IEEE Computer Society 2008: Proceedings of the 2008 International Conference on BioMedical Engineering and Informatics, Sanya, Hainan, China, Vol. 2, pp. 867-871.

Synek, V., (2001). Calibration lines passing through the origin with errors in both axes. *Accred. Qual. Assur.* 6(8), 360-367.

Synek, V., Subrt, P., Marecek, J., (2000). Uncertainties of mercury determinations in biological materials using an atomic absorption spectrometer – AMA 254. *Accred. Qual. Assur.* 5(2), 58-66.

Tan, H.S., Jones, W.E., (1989). Fitting of a straight line when both variables contain errors. Application to the Beer-Lambert law. *J. Chem. Educ.* 66(8), 650-651.

Turner, M.E., (1960). Straight line regression through the origin. *Biometrics* 16(3), 483-485.

Uyar, B., Erdem, O., (1990). Regression procedures in SAS problems?. *Am. Stat.* 44(4), 296-301.

Valentine, T.J., (1971). Regression through origin. *Am. Stat.* 25(5), 58-59.

van Zoonen, P., Hoogerbrugge, R., Gort, S.M., van de Wiel, H.J., van't Klooster, H.A., (1999). Some practical examples of method validation in the analytical laboratory. *Trends Anal. Chem.* 18(9-10), 584-593.

Waters, E.K., Furlon, M.J., Benke, K.K., Grove, J.R., Hamilton, A.J., (2014). Iwao's patchiness regression through the origin: biological importance and efficiency of sampling applications. *Popul. Ecol.* 56(2), 393-399.

Willett, J.B., Singer, J.D., (1988). Another cautionary note about R^2, it use in weighted least squares regression. *Am. Stat.* 42(3), 236-238.

Williamson, J.H., (1968). Least squares fitting of a straight line. *Can. J. Phys*. 46 (16), 1845-1847.

Winsor, C.P., (1946). Which regression. *Biometr. Bull.* 2(6), 101-109.

York, D., (1969). Least squares fitting of a straight line with correlated errors. *Earth Planet. Sci. Lett.* 5, 320-324.

In: Linear Regression
Editor: Vera L. Beck
ISBN: 978-1-53611-992-3
© 2017 Nova Science Publishers, Inc.

Chapter 3

LINEAR REGRESSION FOR INTERVAL-VALUED DATA IN $\mathcal{K}_C(\mathbb{R})$

Yan Sun**and Chunyang Li

Department of Mathematics & Statistics
Utah State University, Logan, UT, US

Abstract

In the recent scientific research, data are increasingly taking on new formats such as sets, lists, and histograms. Among these, a particular type that is frequently encountered is interval-valued data, which refers to a collection of observations in the form of intervals. Some examples include daily [min, max] temperature, [low, high] elevation of a geographical region, and the range of a group of individual observations. Linear regression as a fundamental tool of statistical analysis has been increasingly investigated for extensions to accommodate interval-valued data. Various models and methods have been proposed and studied in the last decades. However, issues such as interpretability and computational feasibility still remain. Especially, a commonly accepted mathematical foundation is largely underdeveloped, compared to the demand of applications.

*Corresponding Author: yan.sun@usu.edu

In this chapter, we focus on linear regression for interval-valued data within the framework of random sets, and propose a new model that generalizes a series of existing ones. By proposing our model, we continue to build up the theoretical framework that deeply understands the existing models and facilitates future developments. In particular, we establish important properties of the model in the space of compact convex subsets of R, analogous to those for the classical linear regression. Additionally, we carry out theoretical investigations into the least squares estimation that is widely used in the literature. It is shown that the least squares estimator is asymptotically unbiased. A simulation study is presented that supports our theorems, and an application to a climate data set is demonstrated.

Keywords: linear regression; random interval, metric space, coefficient of determination, least squares, asymptotic unbiasedness

1. Introduction

Linear regression for interval-valued data has been attracting increasing interests among researchers. See [10], [20], [12, 13], [23], [8], [5], [14], [26, 27], [6], [9], for a partial list of references. However, issues such as interpretability and computational feasibility still remain. Especially, a commonly accepted mathematical foundation is largely underdeveloped, compared to its demand of applications. By proposing our new model, we continue to build up the theoretical framework that deeply understands the existing models and facilitates future developments.

In the statistics literature, the interval-valued data analysis is most often studied under the framework of random sets, which includes random intervals as the special (one-dimensional) case. The probability-based theory for random sets has developed since the publication of the seminal book of [24]. See [25] for a relatively complete monograph. To facilitate the presentation of our results, we briefly introduce the basic notations and definitions in the random set theory. Let $(\Omega, \mathcal{L}, P)$ be a probability space. Denote by $\mathcal{K}\left(\mathbb{R}^d\right)$ or $\mathcal{K}$ the collection of all non-empty compact subsets of $\mathbb{R}^d$. In the space $\mathcal{K}$, a linear structure is defined by Minkowski addition and scalar multiplication, i.e.,

$$A+B=\{a+b:a\in A,b\in B\}\quad \lambda A=\{\lambda a:a\in A\},$$

$\forall A,B\in\mathcal{K}$ and $\lambda\in\mathbb{R}$. A natural metric for the space $\mathcal{K}$ is the Hausdorff metric ρ_H, which is defined as

$$\rho_H(A,B)=\max\left(\sup_{a\in A}\rho(a,B),\sup_{b\in B}\rho(b,A)\right),\ \forall A,B\in\mathcal{K},$$

where ρ denotes the Euclidean metric. A random compact set is a Borel measurable function $A:\Omega\to\mathcal{K}$, $\mathcal{K}$ being equipped with the Borel σ-algebra induced by the Hausdorff metric. For each $X\in\mathcal{K}\left(\mathbb{R}^d\right)$, the function defined on the unit sphere S^{d-1}:

$$s_X(u)=\sup_{x\in X}\langle u,x\rangle,\ \ \forall u\in S^{d-1}$$

is called the support function of X. If $A(\omega)$ is convex almost surely, then A is called a random compact convex set. (See [25], p.21, p.102.) The collection of all compact convex subsets of $\mathbb{R}^d$ is denoted by $\mathcal{K}_{\mathcal{C}}\left(\mathbb{R}^d\right)$ or $\mathcal{K}_{\mathcal{C}}$. When $d=1$, the corresponding $\mathcal{K}_{\mathcal{C}}$ contains all the non-empty bounded closed intervals in $\mathbb{R}$. A measurable function $X:\Omega\to\mathcal{K}_{\mathcal{C}}(\mathbb{R})$ is called a random interval. Much of the random sets theory has focused on compact convex sets. Let $\mathcal{S}$ be the space of support functions of all non-empty compact convex subsets in $\mathcal{K}_{\mathcal{C}}$. Then, $\mathcal{S}$ is a Banach space equipped with the L_2 metric

$$\|s_X(u)\|_2=\left[d\int_{S^{d-1}}|s_X(u)|^2\mu(\mathrm{d}u)\right]^{\frac{1}{2}},$$

where μ is the normalized Lebesgue measure on S^{d-1}. According to the embedding theorems (see [28], [15]), $\mathcal{K}_{\mathcal{C}}$ can be embedded isometrically into the Banach space $C(S)$ of continuous functions on S^{d-1}, and $\mathcal{S}$ is the image of $\mathcal{K}_{\mathcal{C}}$ into $C(S)$. Therefore, $\delta(X,Y):=\|s_X-s_Y\|_2$, $\forall X,Y\in\mathcal{K}_{\mathcal{C}}$, defines a metric on $\mathcal{K}_{\mathcal{C}}$. Particularly, let

$$X=[\underline{X},\overline{X}]=[X^c-X^r,X^c+X^r]$$

be an bounded closed interval with center X^c and radius X^r, or lower bound $\underline{X}$ and upper bound $\overline{X}$, respectively. Then, the δ-metric of X is

$$\|X\|_2=\|s_X(u)\|_2=\frac{1}{2}\left(\underline{X}^2+\overline{X}^2\right)=(X^c)^2+(X^r)^2,$$

and the δ-distance between two intervals X and Y is

$$\begin{aligned}\delta(X,Y) &= \left[\frac{1}{2}\left(\underline{X}-\underline{Y}\right)^2+\frac{1}{2}\left(\overline{X}-\overline{Y}\right)^2\right]^{\frac{1}{2}} \\ &= \left[\left(X^c-Y^c\right)^2+\left(X^r-Y^r\right)^2\right]^{\frac{1}{2}}.\end{aligned}$$

Investigation of linear regression in $\mathcal{K}_{\mathcal{C}}(\mathbb{R})$ began with [10] developing a least squares fitting of compact set-valued data and considering the interval-valued input and output as a special case. Precisely, he gave analytical solutions to the real-valued numbers a and b under different circumstances such that $\delta(Y,aX+b)$ is minimized on the data. The pioneer idea of [10] was further studied in [11, 12], where the δ-metric was extended to a more general metric called W-metric originally proposed by [20]. The advantage of the W-metric lies in the flexibility to assign weights to the radius and midpoints in calculating the distance between intervals. So far the literature had been focusing on finding the affine transformation $Y=aX+b$ that best fits the data, but the data are not assumed to fulfill such a transformation. A probabilistic model along this direction kept missing until [13], and simultaneously [14], proposed the same simple linear regression model for the first time. The model essentially takes on the form of

$$Y_i = aX_i+b+\varepsilon_i, \tag{1}$$

with $a,b\in\mathbb{R}$ and $E(\varepsilon_i)=[-c,c], c\in\mathbb{R}$. This can be written equivalently as

$$\begin{aligned}Y_i^c &= aX_i^c+b+\varepsilon_i^c, \\ Y_i^r &= |a|X_i^r+c+\varepsilon_i^r.\end{aligned}$$

It leads to the following equation that shows linearity in $\mathcal{K}_{\mathcal{C}}$:

$$\delta\left(\hat{Y}_i,\hat{Y}_j\right)=|a|\delta\left(X_i,X_j\right). \tag{2}$$

Some advances have been made regarding this model and the associated estimators. [13] derived least squares estimators for the model parameters and examined them from a theoretical perspective. [14] established a test of linear independence for interval-valued data. However, many problems still remain open such as biases and asymptotic distributions, as anticipated in [13]. This

chapter presents a continuous development addressing some issues and open problems in the direction of model (1).

We point out that, in a separate framework, linear regression models for interval-valued data have been studied in $\mathbb{R}^2$ by treating the intervals essentially as bivariate vectors. Examples belonging to this category include the center method by [3], the MinMax method by [4], the (constrained) center and range method by [26, 27], and the model M by [6]. Although the bivariate representation of an interval could result in loss of geometric information (e.g., equation (2) does not hold anymore), this type of models generally has better flexibility and easier inferences, and therefore are preferred in some practical situations. We emphasize that the purpose of the chapter is not to compare models from the two domains, but to focus on and provide insights into model developments in $\mathcal{K}_C(\mathbb{R})$.

Our contributions in this chapter are three fold. First, we relax the restriction of model (1) that the Hukuhara difference $Y \ominus (aX+b)$ must exist (see [16]) and generalize the univariate model to the multiple case. We also give analytical least squares (LS) solutions to the model parameters. Second, we show that our model and LS estimation together accommodate a decomposition of the sums of squares in $\mathcal{K}_C$ analogous to that of the classical linear regression. Third, we derive explicit formulas of the LS estimates for the univariate model, which exist with probability going to one. The LS estimates are further shown to be asymptotically unbiased. A simulation study is carried out to validate our theoretical findings. Finally, we apply our model to a climate data set to illustrate the applicability of our model.

The rest of the chapter is organized as follows: Section 2 formally introduces our model and the associated LS estimators. Then, the sums of squares and coefficient of determination in $\mathcal{K}_C$ are defined and discussed. Section 3 presents the theoretical properties of the LS estimates for the univariate model. The simulation study is reported in Section 4, and the real data application is presented in Section 5. We give concluding remarks in Section 6. Technical proofs and useful lemmas are deferred to the Appendices.

2. The Proposed Model

2.1 Model Specification

We consider an extension of model (1) to the form

$$\delta(Y_i, aX_i + b) = \|\varepsilon_i\|_2, \tag{3}$$

where $E[\varepsilon_i] = [-c, c]$, $c > 0$. It is equivalently expressed as

$$\begin{cases} Y_i = aX_i + b + \varepsilon_i, \text{ if } Y_i \ominus (aX_i + b) \text{ exists;} \\ Y_i + \varepsilon_i = aX_i + b, \text{ if otherwise } (aX_i + b) \ominus Y_i \text{ exists.} \end{cases} \tag{4}$$

This leads to the following center-radius specification

$$\begin{aligned} Y_i^c &= aX_i^c + b \pm \varepsilon_i^c, \\ Y_i^r &= |a| X_i^r \pm \varepsilon_i^r, \end{aligned}$$

where $E(\varepsilon_i^c) = 0$, $E(\varepsilon_i^r) = c > 0$, and the signs "$\pm$" correspond to the two cases in (4). Define

$$\begin{cases} \lambda_i = \varepsilon_i^c, \ \eta_i = \varepsilon_i^r, & \text{if } Y_i \ominus (aX_i + b) \text{ exists;} \\ \lambda_i = -\varepsilon_i^c, \ \eta_i = -\varepsilon_i^r, & \text{if otherwise } (aX_i + b) \ominus Y_i \text{ exists.} \end{cases} \tag{5}$$

Our model is specified as

$$Y_i^c = aX_i^c + b + \lambda_i, \tag{6}$$

$$Y_i^r = |a| X_i^r + \eta_i, \tag{7}$$

where $\mathrm{E}(\lambda_i) = 0$, $\mathrm{E}(\eta_i) = \mu \in [-c, c]$, $\mathrm{Var}(\lambda_i) = \sigma_\lambda^2 > 0$, and $\mathrm{Var}(\eta_i) = \sigma_\eta^2 > 0$.

To model the outcome intervals $Y_i = \left[\underline{Y_i}, \overline{Y_i}\right]$ by p interval-valued predictors $X_{j,i} = \left[\underline{X_{j,i}}, \overline{X_{j,i}}\right]$, $i = 1, \cdots, n$; $j = 1, \cdots, p$, we consider the multivariate extension of (3):

$$\delta\left(Y_i, b + \sum_{j=1}^{p} a_j X_{j,i}\right) = \|\varepsilon_i\|_2, \tag{8}$$

which leads to the following center-radius specification

$$Y_i^c = b + \sum_{j=1}^{p} a_j X_{j,i}^c + \lambda_i, \tag{9}$$

$$Y_i^r = \sum_{j=1}^{p} |a_j| X_{j,i}^r + \eta_i. \tag{10}$$

where $\mathrm{E}(\lambda_i) = 0$, $\mathrm{E}(\eta_i) = \mu \in [-c, c]$, $\mathrm{Var}(\lambda_i) = \sigma_\lambda^2$, and $\mathrm{Var}(\eta_i) = \sigma_\eta^2$. We have assumed λ_i and η_i are independent in this chapter to simplify the presentation. The model that includes a covariance between λ_i and η_i can be implemented without much extra difficulty.

2.2 Least Squares Estimate (LSE)

Least squares method is widely used in the literature to estimate the interval-valued regression coefficients ([10], [20], [12]). It minimizes $\delta(Y, \mathrm{E}(Y|X))$ on the data with respect to the parameters. Denote

$$\hat{Y}_i^c = \mathrm{E}(Y_i^c|X_i) = b + \sum_{j=1}^{p} a_j X_{j,i}^c, \tag{11}$$

$$\hat{Y}_i^r = \mathrm{E}(Y_i^r|X_i) = \mu + \sum_{j=1}^{p} |a_j| X_{j,i}^r. \tag{12}$$

Then the sum of squared δ-distance between Y_i and $\hat{Y}_i$ is written as

$$\begin{aligned} L &= \sum_{i=1}^{n} \delta^2 \left[\mathrm{E}(Y_i|X_i), Y_i \right] \\ &= \sum_{i=1}^{n} \left[\left(b + \sum_{j=1}^{p} a_j X_{j,i}^c - Y_i^c \right)^2 + \left(\sum_{j=1}^{p} |a_j| X_{j,i}^r + \mu - Y_i^r \right)^2 \right]. \end{aligned}$$

Therefore, the LSE of $\{\mu, b, a_j, j = 1, \cdots, p\}$ is defined as

$$\{\hat{\mu}, \hat{b}, \hat{a}_j, j = 1, \cdots, p\} = \arg\min \left\{ \frac{1}{n} L(\mu, b, a_j, j = 1, \cdots, p) \right\}. \tag{13}$$

Let

$$\begin{aligned} S\left(X_j^c, X_k^c\right) &= \frac{1}{n}\sum_{i=1}^{n} X_{j,i}^c X_{k,i}^c - \left(\frac{1}{n}\sum_{i=1}^{n} X_{j,i}^c\right)\left(\frac{1}{n}\sum_{i=1}^{n} X_{k,i}^c\right), \\ S\left(X_j^r, X_k^r\right) &= \frac{1}{n}\sum_{i=1}^{n} X_{j,i}^r X_{k,i}^r - \left(\frac{1}{n}\sum_{i=1}^{n} X_{j,i}^r\right)\left(\frac{1}{n}\sum_{i=1}^{n} X_{k,i}^r\right), \end{aligned}$$

be the sample covariances of the centers and radii of X_j and X_k, respectively. Especially, when $k = j$, we denote by $S^2\left(X_j^c\right)$ and $S^2\left(X_j^r\right)$ the corresponding sample variances. In addition, define

$$\begin{aligned} S\left(X_j^c, Y^c\right) &= \frac{1}{n}\sum_{i=1}^{n} X_{j,i}^c Y^c - \left(\frac{1}{n}\sum_{i=1}^{n} X_{j,i}^c\right)\left(\frac{1}{n}\sum_{i=1}^{n} Y^c\right), \\ S\left(X_j^r, Y^r\right) &= \frac{1}{n}\sum_{i=1}^{n} X_{j,i}^r Y^r - \left(\frac{1}{n}\sum_{i=1}^{n} X_{j,i}^r\right)\left(\frac{1}{n}\sum_{i=1}^{n} Y^r\right), \end{aligned}$$

as the sample covariances of the centers and radii of X_j and Y, respectively. Then, the minimization problem (13) is solved in the following proposition.

Proposition 1. *The least squares estimates of the regression coefficients* $\left\{\hat{a}_j\right\}_{j=1}^{p}$, *if they exist, are solution of the equation system:*

$$\begin{aligned} &\sum_{j=1}^{p} a_j S\left(X_j^c, X_k^c\right) + sgn\left(a_k\right)\sum_{j=1}^{p} |a_j| S\left(X_j^r, X_k^r\right) \\ = \;& S\left(X_k^c, Y^c\right) + sgn\left(a_k\right) S\left(X_k^r, Y^r\right), \;\; k = 1, \cdots, p. \end{aligned} \tag{14}$$

And then, $\left\{\hat{b}, \hat{\mu}\right\}$ *are given by*

$$\hat{b} = \overline{Y^c} - \sum_{j=1}^{p} \hat{a}_j \overline{X_j^c}, \tag{15}$$

$$\hat{\mu} = \overline{Y^r} - \sum_{j=1}^{p} |\hat{a}_j| \overline{X_j^r}. \tag{16}$$

2.3 Sums of Squares and Coefficient of Determination

The variance of a compact convex random set X in $\mathbb{R}^d$ is defined via its support function as

$$\mathrm{Var}(X) = \mathrm{E}\delta^2(X, \mathrm{E}X),$$

where the expectation is defined by Aumann integral (see [2], [1]) as

$$\mathrm{E}X = \{\mathrm{E}\xi : \xi \in X \text{ almost surely}\}.$$

See [18, 19]. For the case $d = 1$, it is shown by straightforward calculations that

$$\begin{aligned}\mathrm{E}X &= [\mathrm{E}\underline{X}, \mathrm{E}\overline{X}],\\ \mathrm{Var}(X) &= \mathrm{Var}(X^c) + \mathrm{Var}(X^r).\end{aligned}$$

This leads us to define the sums of squares in $\mathcal{K}_C(\mathbb{R})$ to measure the variability of interval-valued data. A definition of the coefficient of determination R^2 in $\mathcal{K}_C(\mathbb{R})$ follows immediately, which produces a measure of goodness-of-fit.

Definition 1. *The total sum of squares (SST) in $\mathcal{K}_C$ is defined as*

$$SST = \sum_{i=1}^{n} \left[\left(Y_i^c - \overline{Y^c}\right)^2 + \left(Y_i^r - \overline{Y^r}\right)^2 \right]. \tag{17}$$

Definition 2. *The explained sum of squares (SSE) in $\mathcal{K}_C$ is defined as*

$$SSE = \sum_{i=1}^{n} \left[\left(\hat{Y}_i^c - \overline{Y^c}\right)^2 + \left(\hat{Y}_i^r - \overline{Y^r}\right)^2 \right]. \tag{18}$$

Definition 3. *The residual sum of squares (SSR) in $\mathcal{K}_C$ is defined as*

$$SSR = \sum_{i=1}^{n} \left[\left(Y_i^c - \hat{Y}_i^c\right)^2 + \left(Y_i^r - \hat{Y}_i^r\right)^2 \right]. \tag{19}$$

Definition 4. *The coefficient of determination (R^2) in $\mathcal{K}_C$ is defined as*

$$R^2 = 1 - \frac{SSR}{SST}, \tag{20}$$

where SST and SSR are defined in (17) and (19), respectively.

Analogous to the classical theory of linear regression, our model (9)-(10) together with the LS estimates (13) accommodates the partition of SST into SSE and SSR. As a result, the coefficient of determination (R^2) can also be calculated as the ratio of SSE and SST. The partition has a series of important implications of the underlying model, one of which being that the residual $Y \ominus \hat{Y} / \hat{Y} \ominus Y$ and the predictor $\hat{Y}$ are empirically uncorrelated in $(\mathcal{K}_C, \delta)$.

Theorem 1. *Assume model (9)-(10). Let Y_i^c and Y_i^r in (11)-(12) be calculated according to the LS estimates $\left\{\hat{\mu}, \hat{b}, \hat{a}_j, j = 1, \cdots, p\right\}$ in (13). Then,*

$$SST = SSE + SSR.$$

It follows that the coefficient of determination in $\mathcal{K}_C$ is equivalent to

$$R^2 = SSE/SST.$$

2.4 Positive Restriction and Goodness-of-fit

It is possible to get negative values of $\hat{Y}_i^r$ by its definition (12). That is, the model implied outcome could be outside $\mathcal{K}_C(\mathbb{R})$. This is an inevitable drawback to force a linear model in the nonlinear space $\mathcal{K}_C(\mathbb{R})$ (e.g., there is no inverse of addition). Theoretically, this phenomenon is closely related to the goodness-of-fit of the linear regression model. Theorem 2 gives an upper bound of how often the model predicts outcomes outside of $\mathcal{K}_C(\mathbb{R})$. For a model that largely explains the variability of Y^r, σ_η^2 should be very small and so is this bound. Otherwise, the upper bound probability could grow large if most of the variability of Y^r lies in the random error. In practice, for model inferences, the negative values of $\hat{Y}_i^r$ can be rounded to 0, which always improves on the predicting accuracy since Y_i^r is non-negative.

Theorem 2. *Consider model (9)-(10). Let $\hat{Y}_i$ be defined in (11)-(12). Then,*

$$P\left(\hat{Y}_i^r < 0\right) \leq \frac{E\left(Y_i^r - \hat{Y}_i^r\right)^2}{\left(Y_i^r\right)^2} = \frac{\sigma_\eta^2}{\left(Y_i^r\right)^2}.$$

3. Properties of LSE

In this section, we study the theoretical properties of the LSE for the univariate model (6)-(7). Applying Proposition 1 to the case $p = 1$, we obtain the two sets of half-space solutions, corresponding to $a \geq 0$ and $a < 0$, respectively, as follows:

$$a^+ = \frac{S(X^c,Y^c)+S(X^r,Y^r)}{S^2(X^c)+S^2(X^r)}, \tag{21}$$

$$b^+ = \overline{Y^c} - a^+\overline{X^c}, \tag{22}$$

$$\mu^+ = \overline{Y^r} - |a^+|\overline{X^r}; \tag{23}$$

and

$$a^- = \frac{S(X^c,Y^c)-S(X^r,Y^r)}{S^2(X^c)+S^2(X^r)}, \tag{24}$$

$$b^- = \overline{Y^c} - a^-\overline{X^c}, \tag{25}$$

$$\mu^- = \overline{Y^r} - |a^-|\overline{X^r}. \tag{26}$$

The final formula for the LS estimates falls in three categories. In the first, there is one and only one set of existing solution, which is defined as the LSE. In the second, both sets of solutions exist, and the LSE is the one that minimizes L. In the third situation, neither solution exists, but this only happens with probability going to 0. We conclude these findings in the following Theorem.

Theorem 3. *Assume model (6)-(7). Let $\{\hat{a},\hat{b},\hat{\mu}\}$ be the least squares solution defined in (13). If $|S(X^c,Y^c)| > |S(X^r,Y^r)|$, then there exists one and only one half-space solution. More specifically,*

i. *if in addition $S(X^c,Y^c) > 0$, then the LS solution is given by*

$$\{\hat{a},\hat{b},\hat{\mu}\} = \{a^+,b^+,\mu^+\};$$

ii. *if instead $S(X^c,Y^c) < 0$, then the LS solution is given by*

$$\{\hat{a},\hat{b},\hat{\mu}\} = \{a^-,b^-,\mu^-\}.$$

Otherwise, $|S(X^c,Y^c)| < |S(X^r,Y^r)|$, and then either both of the half-space solutions exist, or neither one exists. In particular,

iii. *if in addition $S(X^r,Y^r) > 0$, then both of the half-space solutions exist, and*

$$\left\{\hat{a},\hat{b},\hat{\mu}\right\} = \arg\min_{\{\{a^+,b^+,\mu^+\},\{a^-,b^-,\mu^-\}\}} \{L(a,b,\mu)\};$$

iv. *if instead $S(X^r,Y^r) < 0$, then the LS solution does not exist, but this happens with probability converging to 0.*

Unlike the classical linear regression, LS estimates for the model (6)-(7) are biased. We calculate the biases explicitly in Proposition 2, which are shown to converge to zero as the sample size increases to infinity. Therefore, the LS estimates are asymptotically unbiased.

Proposition 2. *Let $\left\{\hat{a},\hat{b},\hat{\mu}\right\}$ be the least squares solution in Theorem 3. Then,*

$$E(\hat{a}-a) = -\frac{2aS^2(X^r)}{S^2(X^c)+S^2(X^r)}\left[P(\hat{a}=a^-)I_{\{a\geq 0\}} + P(\hat{a}=a^+)I_{\{a<0\}}\right],$$

$$E(|\hat{a}|-|a|) = -\frac{2|a|S^2(X^c)}{S^2(X^c)+S^2(X^r)}\left[P(\hat{a}=a^-)I_{\{a\geq 0\}} + P(\hat{a}=a^+)I_{\{a<0\}}\right].$$

Theorem 4. *Consider model (6)-(7). Assume $S^2(X^c) = O(1)$ and $S^2(X^r) = O(1)$. Then, the least squares solution $\left\{\hat{a},\hat{b},\hat{\mu}\right\}$ in Theorem 3 is asymptotically unbiased, i.e.*

$$E\begin{bmatrix}\hat{a}\\ \hat{b}\\ \hat{\mu}\end{bmatrix} \to \begin{bmatrix}a\\ b\\ \mu\end{bmatrix},$$

as $n \to \infty$.

4. Simulation

We carry out a systematic simulation study to examine the empirical performance of the least squares method proposed in this chapter. First, we consider the following three models:

- Model 1: $a = 2$, $b = 5$, $\mu = 0.5$, $\sigma_\eta = 0.3$, $\sigma_\lambda = 2$;
- Model 2: $a = -2$, $b = 5$, $\mu = 0.5$, $\sigma_\eta = 0.3$, $\sigma_\lambda = 3$;
- Model 3: $a = 2$, $b = 5$, $\mu = -0.5$, $\sigma_\eta = 0.3$, $\sigma_\lambda = 2$;

where data show a positive correlation, a negative correlation, and a positive correlation with a negative μ, respectively. A simulated dataset from each model is shown in Figure 1, along with its fitted regression line.

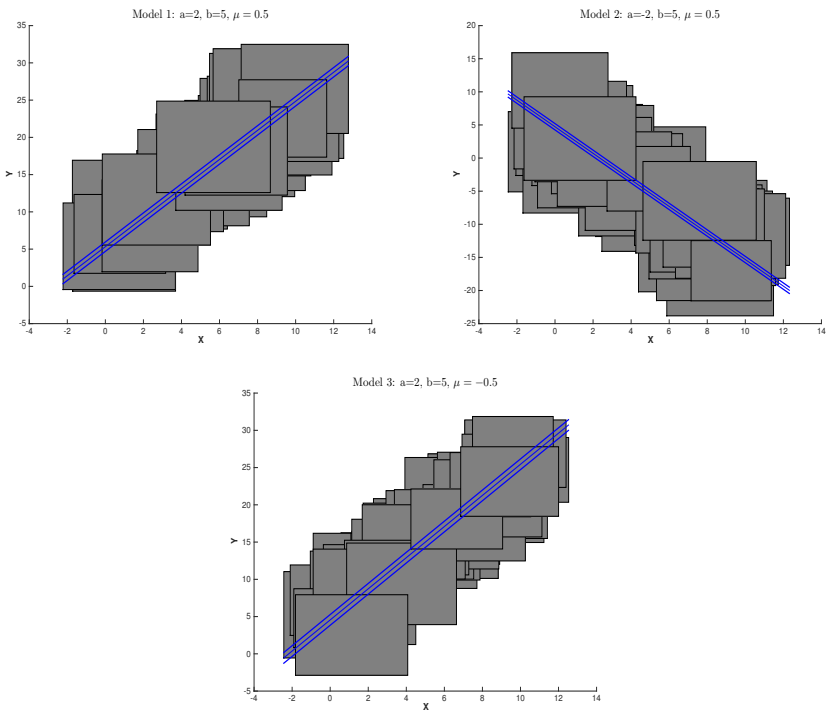

Figure 1: Plots of simulated datasets from models 1, 2, and 3, each with sample size $n = 50$. The solid line denotes the regression line $y = \hat{a}x + \hat{b}$, and the two dashed lines denote the two accompanying lines $y = \hat{a}x + \hat{b} \pm \hat{\mu}$.

To investigate the asymptotic behavior of the LS estimates, we repeat the process of data generation and parameter estimation 1000 times independently using sample size $n = 20, 50, 100$ for all the three models. The resulting 1000 independent sets of parameter estimates for each model/sample size are evaluated by their mean absolute error (MAE) and mean error (ME). The numerical results are summarized in Table 1. Consistent with Proposition 2, $\hat{a}$ tends to underestimate a when $a > 0$ and overestimate a when $a < 0$. This bias also causes a positive and negative bias in $\hat{b}$, when $a > 0$ and $a < 0$, respectively. Similarly, a positive bias in $\hat{\mu}$ is induced by the negative bias in $|\hat{a}|$. All the biases diminish to 0 as the sample size increases to infinity, which confirms our finding in Theorems 4.

Table 1. Evaluation of Parameter Estimation

	n	MAE	ME	MAE	ME	MAE	ME
		a=2		**b=5**		$\mu = 0.5$	
Model 1	20	0.1449	-0.0921	0.8083	0.445	0.3655	0.2304
	50	0.0848	-0.0411	0.4899	0.2141	0.214	0.1011
	100	0.0562	-0.0171	0.3151	0.0872	0.142	0.041
		a=-2		**b=5**		$\mu = 0.5$	
Model 2	20	0.2011	0.103	1.1389	-0.5071	0.5067	0.2578
	50	0.1205	0.0336	0.6973	-0.1774	0.3038	0.0807
	100	0.0842	0.0185	0.4814	-0.0865	0.2118	0.0465
		a=2		**b=5**		$\mu = -0.5$	
Model 3	20	0.1488	-0.1047	0.8143	0.495	0.3785	0.262
	50	0.0836	-0.0412	0.4703	0.2119	0.2108	0.1015
	100	0.0579	-0.0187	0.3321	0.098	0.1453	0.0464

Next, we compare our model to CCRM, a typical bivariate type of model from the literature. As we discussed in the introduction, these two models are developed for different purposes and are generally not comparable. We include a comparison in the simulation study to better evaluate the performances of

our model, with CCRM providing a baseline of converging rate and predicting accuracy. From Model 1, 2, 3, respectively, we simulate 1000 independent samples with size $n = 20, 50, 100$. Then, each sample is randomly split into a training set (80%) and a validation set (20%). The two models are evaluated by their sample variance adjusted mean squared errors (AMSE's) on the validation set, which are defined as

$$\text{AMSE(center)} = \frac{\sum_{i=1}^{m} \left(Y_i^c - \hat{Y}_i^c\right)^2}{\sum_{i=1}^{m} \left(Y_i^c - \overline{Y}_i^c\right)^2},$$
$$\text{AMSE(radius)} = \frac{\sum_{i=1}^{m} \left(Y_i^r - \hat{Y}_i^r\right)^2}{\sum_{i=1}^{m} \left(Y_i^r - \overline{Y}_i^r\right)^2},$$

and

$$\text{AMSE(average)} = \frac{\text{AMSE(center)} + \text{AMSE(radius)}}{2},$$

where $m = n/5$ is the size of validation set. We use the R function *ccrm* in the iRegression package to implement CCRM. The average result of the 1000 repetitions are summarized in Table 2. For Model 1 and 2, both models have competitive performances. Model 3 has a negative μ, so CCRM is slightly worse than our model due to its positive restriction on μ. To better show this, we continue to consider the following two univariate models and one multivariate model with a much smaller μ:

- Model 4: $a = 3$, $b = 5$, $\mu = -5$, $\sigma_\eta = 0.5$, $\sigma_\lambda = 5$;
- Model 5: $a = -3$, $b = 5$, $\mu = -5$, $\sigma_\eta = 0.5$, $\sigma_\lambda = 5$;
- Model 6: $a_1 = -3$, $a_2 = 2$ $b = 5$, $\mu = -5$, $\sigma_\eta = 0.5$, $\sigma_\lambda = 5$.

A sample of $n = 50$ from each of Model 4 and 5 are plotted in Figure 2. For all of the three models, our model performs significantly better than CCRM.

5. A Real Data Application

In this section, we apply our model to analyze the average temperature data for large US cities, which are provided by National Oceanic and Atmospheric

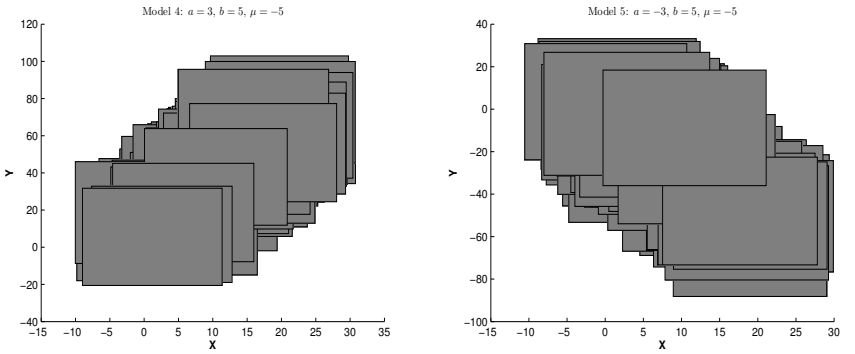

Figure 2: Plots of simulated datasets from models 4 and 5, each with sample size $n = 50$.

Administration (NOAA) and are publicly available. The three data sets we obtained specifically are average temperatures for 51 large US cities in January, April, and July. Each observation contains the averages of minimum and maximum temperatures based on weather data collected from 1981 to 2010 by the NOAA National Climatic Data Center of the United States. July in general is the hottest month in the US. By this analysis, we aim to predict the summer (July) temperatures by those in the winter (January) and spring (April). Figure 3 plots the July temperatures versus those in January and April, respectively.

The parameters are estimated according to (14)-(16) as

$$\hat{a}_1 = -0.4831,\ \hat{a}_2 = 1.1926;$$
$$\hat{b} = 10.2510,\ \hat{\mu} = -3.7071.$$

Denote by T_{Jan}, T_{April}, and T_{July}, the average temperatures in a US city in January, April, and July, respectively. The prediction for T_{July} based on T_{Jan} and T_{April} is given by

$$\hat{T}^c_{July} = 10.2510 - 0.4831T^c_{Jan} + 1.1926T^c_{April}, \quad (27)$$
$$\hat{T}^r_{July} = -3.7071 + 0.4831T^r_{Jan} + 1.1926T^r_{April}. \quad (28)$$

The three sums of squares are calculated to be

$$SST = 663.8627;\ SSE = 495.0874;\ SSR = 168.7753.$$

Table 2. Mean results of AMSE on the validation set based on 1000 independent repetitions

		CCRM			Our Model		
	n	Center	Radius	Average	Center	Radius	Average
Model 1	20	0.1716	0.3134	**0.2425**	0.1772	0.3374	0.2573
	50	0.1181	0.2368	0.1775	0.1116	0.2313	**0.1714**
	100	0.1149	0.2241	0.1695	0.1119	0.2219	**0.1669**
Model 2	20	0.3499	0.3244	**0.3372**	0.3467	0.3294	0.3380
	50	0.2341	0.2356	0.2348	0.2344	0.2318	**0.2331**
	100	0.2263	0.2201	0.2232	0.2203	0.2200	**0.2201**
Model 3	20	0.1708	0.3367	0.2538	0.1687	0.3241	**0.2464**
	50	0.1192	0.2288	0.1740	0.1192	0.2246	**0.1719**
	100	0.1128	0.2196	0.1662	0.1101	0.2190	**0.1646**
Model 4	20	0.3795	0.3499	0.3647	0.1250	0.3691	**0.2470**
	50	0.2802	0.2738	0.2770	0.0867	0.2734	**0.1800**
	100	0.2580	0.2727	0.2653	0.0808	0.2605	**0.1706**
Model 5	20	0.3519	0.3207	0.3363	0.1204	0.3717	**0.2461**
	50	0.2712	0.2799	0.2756	0.0827	0.2681	**0.1754**
	100	0.2558	0.2751	0.2655	0.0800	0.2552	**0.1676**
Model 6	50	0.0622	0.4288	0.2455	0.0661	0.2536	**0.1599**
	100	0.0596	0.3934	0.2265	0.0606	0.2370	**0.1488**
	200	0.0565	0.3838	0.2201	0.0593	0.2344	**0.1469**

Therefore, the coefficient of determination is

$$R^2 = 1 - \frac{SSR}{SST} = \frac{SSE}{SST} = 0.7458.$$

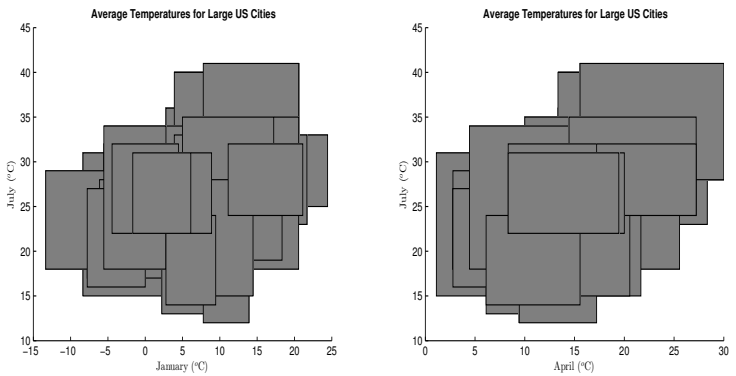

Figure 3: Left: plot of July versus January temperatures. Right: plot of July versus April temperatures.

Finally, the variance parameters can be estimated as

$$\begin{aligned}\hat{\sigma}^2_{\lambda} &= \frac{1}{n-1}\sum_{i=1}^{n}\left(T^c_{July,i}-\hat{T}^c_{July,i}\right)^2 = 2.1708;\\ \hat{\sigma}^2_{\eta} &= \frac{1}{n-1}\sum_{i=1}^{n}\left(T^r_{July,i}-\hat{T}^r_{July,i}\right)^2 = 1.2047.\end{aligned}$$

Thus, by Theorem 2, an upper bound of $P\left(\hat{T}^r_{July,i}<0\right)$ on average is estimated to be

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{\sigma}^2_{\eta}}{\left(T^r_{July,i}\right)^2} = \frac{1.2047}{n}\sum_{i=1}^{n}\frac{1}{\left(T^r_{July,i}\right)^2} = 0.047,$$

which is very small and reasonably ignorable. We calculate $\hat{T}^r_{July,i}$ for the entire sample and all of them are well above 0. So, for this data, although $\hat{\mu}<0$ and it is possible to get negative predicted radius, it in fact never happens because the model has captured most of the variability. The empirical distributions of residuals are shown in Figure 4. Both distributions are centered at 0, with the center residual having a slightly bigger tail.

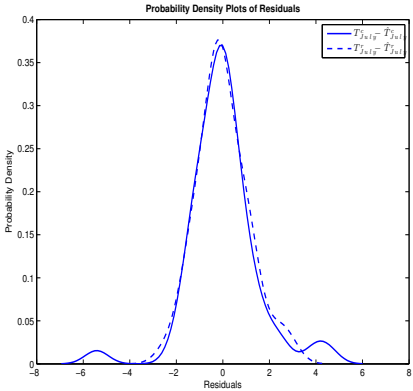

Figure 4: Empirical probability density plots of the residuals for the center and radius.

Conclusion

We have rigorously studied linear regression for interval-valued data in the metric space $(\mathcal{K}_{\mathcal{C}}, \delta)$. The new model we introduces generalizes previous models in the literature so that the Hukuhara difference $Y_i \ominus (aX_i + b)$ needs not exist. Analogous to the classical linear regression, our model together with the LS estimation leads to a partition of the total sum of squares (SSR) into the explained sum of squares (SSE) and the residual sum of squares (SSR) in $(\mathcal{K}_{\mathcal{C}}, \delta)$, which implies that the residual is uncorrelated with the linear predictor in $(\mathcal{K}_{\mathcal{C}}, \delta)$. In addition, we have carried out theoretical investigations into the least squares estimation for the univariate model. It is shown that the LS estimates in $(\mathcal{K}_{\mathcal{C}}, \delta)$ are biased but the biases reduce to zero as the sample size tends to infinity. Therefore, a bias-correction technique for small sample estimation could be a good future topic. The simulation study confirms our theoretical findings and shows that the least squares estimators perform satisfactorily well for moderate sample sizes.

Appendix: Proofs

Proof of Proposition 1

Proof. Differentiating L with respect to μ, b, and a_j, $j = 1, \cdots, p$, respectively, and setting the derivatives to zero, we get

$$\frac{\partial L}{\partial \mu} \propto \sum_{i=1}^{n} \left(\hat{Y}_i^r - Y_i^r\right) = 0, \tag{29}$$

$$\frac{\partial L}{\partial b} \propto \sum_{i=1}^{n} \left(\hat{Y}_i^c - Y_i^c\right) = 0, \tag{30}$$

$$\frac{\partial L}{\partial a_k} \propto \sum_{i=1}^{n} \left(\hat{Y}_i^c - Y_i^c\right) X_{k,i}^c + \sum_{i=1}^{n} \left(\hat{Y}_i^r - Y_i^r\right) sgn\left(a_k\right) X_{k,i}^r = 0, \tag{31}$$

$k = 1, \cdots, p.$

Equations (29)-(30) yield

$$b = \frac{1}{n}\sum_{i=1}^{n} Y_i^c - \frac{1}{n}\sum_{j=1}^{p} a_j \sum_{i=1}^{n} X_{j,i}^c = \overline{Y^c} - \sum_{j=1}^{p} a_j \overline{X_j^c}, \tag{32}$$

$$\mu = \frac{1}{n}\sum_{i=1}^{n} Y_i^r - \frac{1}{n}\sum_{j=1}^{p} |a_j| \sum_{i=1}^{n} X_{j,i}^r = \overline{Y^r} - \sum_{j=1}^{p} |a_j| \overline{X_j^r}. \tag{33}$$

Equations (14) are obtained by plugging (32)-(33) into (31), and equations (15)-(16) follow from (32)-(33). This completes the proof. □

5.1 Proof of Theorem 1

Proof. According to definitions (17)-(19),

$$\begin{aligned} SST &= \sum_{i=1}^{n} \left[\left(Y_i^c - \hat{Y}_i^c + \hat{Y}_i^c - \overline{Y^c}\right)^2 + \left(Y_i^r - \hat{Y}_i^r + \hat{Y}_i^r - \overline{Y^r}\right)^2\right] \\ &= SSE + SSR + 2\sum_{i=1}^{n} \left[\left(Y_i^c - \hat{Y}_i^c\right)\left(\hat{Y}_i^c - \overline{Y^c}\right) + \left(Y_i^r - \hat{Y}_i^r\right)\left(\hat{Y}_i^r - \overline{Y^r}\right)\right] \\ &= SSE + SSR + 2\sum_{i=1}^{n} \left[\left(Y_i^c - \hat{Y}_i^c\right)\hat{Y}_i^c + \left(Y_i^r - \hat{Y}_i^r\right)\hat{Y}_i^r\right]. \end{aligned} \tag{34}$$

The last equation is due to (29)-(30). Further in view of (11)-(12) and (31), we have

$$\begin{aligned}
& \sum_{i=1}^{n}\left[\left(Y_i^c-\hat{Y}_i^c\right)\hat{Y}_i^c+\left(Y_i^r-\hat{Y}_i^r\right)\hat{Y}_i^r\right] \\
= & \sum_{i=1}^{n}\left[\left(Y_i^c-\hat{Y}_i^c\right)\sum_{j=1}^{p}a_jX_{j,i}^c+\left(Y_i^r-\hat{Y}_i^r\right)\sum_{j=1}^{p}|a_j|X_{j,i}^r\right] \\
= & \sum_{j=1}^{p}a_j\sum_{i=1}^{n}\left[\left(Y_i^c-\hat{Y}_i^c\right)X_{j,i}^c+\left(Y_i^r-\hat{Y}_i^r\right)sgn(a_j)X_{j,i}^r\right] \\
= & 0.
\end{aligned}$$

This together with (34) completes the proof. □

5.2 Proof of Theorem 2

Proof. Notice that

$$P\left(\hat{Y}_i^r<0\right)=P\left(\hat{Y}_i^r-Y_i^r<-Y_i^r\right)\leq P\left(|\hat{Y}_i^r-Y_i^r|>Y_i^r\right).$$

An application of Markov's inequality completes the proof. □

5.3 Proof of Theorem 3

Proof. Parts **i**, **ii** and **iii** are obvious from Proposition 1. Part **iv** follows from Lemma 1 in Appendix II. □

5.4 Proof of Proposition 2

Proof. We prove the cases $a\geq 0$ and $a<0$ separately. To simplify notations, we will use $E(\cdot)$ throughout the proof, but the expectation should be interpreted as being conditioned on X.

Case I: $a\geq 0$.

From Lemma 2, we have

$$
\begin{aligned}
& a^{+}-a \\
=& \frac{\sum_{i<j}(X_i^c-X_j^c)(Y_i^c-Y_j^c)+\sum_{i<j}(X_i^r-X_j^r)(Y_i^r-Y_j^r)}{\sum_{i<j}(X_i^c-X_j^c)^2+\sum_{i<j}(X_i^r-X_j^r)^2}-a \\
=& \frac{\sum_{i<j}(X_i^c-X_j^c)\left[(Y_i^c-Y_j^c)-a(X_i^c-X_j^c)\right]+\sum_{i<j}(X_i^r-X_j^r)\left[(Y_i^r-Y_j^r)-a(X_i^r-X_j^r)\right]}{\sum_{i<j}(X_i^c-X_j^c)^2+\sum_{i<j}(X_i^r-X_j^r)^2} \\
=& \frac{\sum_{i<j}(X_i^c-X_j^c)(\lambda_i-\lambda_j)+\sum_{i<j}(X_i^r-X_j^r)(\eta_i-\eta_j)}{\sum_{i<j}(X_i^c-X_j^c)^2+\sum_{i<j}(X_i^r-X_j^r)^2}.
\end{aligned}
$$

This immediately yields

$$E\left(a^{+}-a\right)=0. \tag{35}$$

Similarly,

$$a^{-}-a=\frac{\sum_{i<j}(X_i^c-X_j^c)(\lambda_i-\lambda_j)-\sum_{i<j}(X_i^r-X_j^r)[2a(X_i^r-X_j^r)+(\eta_i-\eta_j)]}{\sum_{i<j}(X_i^c-X_j^c)^2+\sum_{i<j}(X_i^r-X_j^r)^2},$$

and consequently,

$$E\left(a^{-}-a\right)=-\frac{2aS^2(X^r)}{S^2(X^c)+S^2(X^r)}. \tag{36}$$

Notice now

$$\begin{aligned}
E(\hat{a}-a) &= E(\hat{a}-a)I_{\{\hat{a}=a^+\}} + E(\hat{a}-a)I_{\{\hat{a}=a^-\}} \\
&= \int_{\{\hat{a}=a^+\}} (\hat{a}-a)\mathrm{d}\mathbb{P} + \int_{\{\hat{a}=a^-\}} (\hat{a}-a)\mathrm{d}\mathbb{P} \\
&= \int_{\{\hat{a}=a^+\}} (a^+-a)\mathrm{d}\mathbb{P} + \int_{\{\hat{a}=a^-\}} (a^--a)\mathrm{d}\mathbb{P} \\
&= \left[\int_{\{\hat{a}=a^+\}} (a^+-a)\mathrm{d}\mathbb{P} + \int_{\{\hat{a}=a^-\}} (a^+-a)\mathrm{d}\mathbb{P}\right] \qquad (37) \\
&\quad - \left[\int_{\{\hat{a}=a^-\}} (a^+-a)\mathrm{d}\mathbb{P} + \int_{\{\hat{a}=a^-\}} (a^--a)\mathrm{d}\mathbb{P}\right] \\
&= E\left(a^+-a\right) - \int_{\{\hat{a}=a^-\}} (a^+-a^-)\mathrm{d}\mathbb{P} \\
&= -E(a^+-a^-)I_{\{\hat{a}=a^-\}}. \qquad (38)
\end{aligned}$$

Here, equation (38) is due to (35). Recall that

$$\begin{aligned}
a^+-a^- &= \frac{2\sum_{i<j}(X_i^r-X_j^r)(Y_i^r-Y_j^r)}{\sum_{i<j}(X_i^c-X_j^c)^2+\sum_{i<j}(X_i^r-X_j^r)^2} \\
&= \frac{2\sum_{i<j}(X_i^r-X_j^r)\left[a(X_i^r-X_j^r)+(\eta_i-\eta_j)\right]}{\sum_{i<j}(X_i^c-X_j^c)^2+\sum_{i<j}(X_i^r-X_j^r)^2}, \qquad (39)
\end{aligned}$$

since $a \geq 0$. Therefore,

$$\begin{aligned}
E(\hat{a}-a) &= -E\left\{\frac{2\sum_{i<j}(X_i^r-X_j^r)\left[a(X_i^r-X_j^r)+(\eta_i-\eta_j)\right]}{\sum_{i<j}(X_i^c-X_j^c)^2+\sum_{i<j}(X_i^r-X_j^r)^2}\right\} I_{\{\hat{a}=a^-\}} \\
&= -\frac{2\sum_{i<j}\left[|a|(X_i^r-X_j^r)^2P(\hat{a}=a^-)+(X_i^r-X_j^r)E(\eta_i-\eta_j)I_{\{\hat{a}=a^-\}}\right]}{\sum_{i<j}(X_i^c-X_j^c)^2+\sum_{i<j}(X_i^r-X_j^r)^2} \\
&= -\frac{2\sum_{i<j}(X_i^r-X_j^r)^2P(\hat{a}=a^-)}{\sum_{i<j}(X_i^c-X_j^c)^2+\sum_{i<j}(X_i^r-X_j^r)^2} \\
&= -\frac{2aS^2(X^r)}{S^2(X^c)+S^2(X^r)}P(\hat{a}=a^-). \qquad (40)
\end{aligned}$$

Similar to the preceding arguments,

$$\begin{aligned} E(|\hat{a}|-|a|) &= E(|\hat{a}|-a) = E(|\hat{a}|-a)I_{\{\hat{a}=a^+\}} + E(|\hat{a}|-a)I_{\{\hat{a}=a^-\}} \\ &= \int_{\{\hat{a}=a^+\}} (a^+ - a)\mathrm{d}\mathbb{P} + \int_{\{\hat{a}=a^-\}} (-a^- - a)\mathrm{d}\mathbb{P} \\ &= E(a^+ - a) - \int_{\{\hat{a}=a^-\}} (a^+ - a)\mathrm{d}\mathbb{P} - \int_{\{\hat{a}=a^-\}} (-a^- + a)\mathrm{d}\mathbb{P} \\ &= -E(a^+ + a^-)I\{\hat{a} = a^-\}. \end{aligned}$$

Recall again that

$$a^+ + a^- = \frac{2\sum_{i<j}(X_i^c - X_j^c)\left[a(X_i^c - X_j^c) + (\lambda_i - \lambda_j)\right]}{S^2(X^c) + S^2(X^r)}. \tag{41}$$

It follows that

$$E(|\hat{a}|-|a|) = -\frac{2aS^2(X^c)}{S^2(X^c)+S^2(X^r)}P(\hat{a} = a^-). \tag{42}$$

Case II: $a < 0$

In this case, we have

$$a^+ - a = \frac{\sum_{i<j}(X_i^c - X_j^c)(\lambda_i - \lambda_j) + \sum_{i<j}(X_i^r - X_j^r)[-2a(X_i^r - X_j^r) + (\eta_i - \eta_j)]}{S^2(X^c)+S^2(X^r)},$$
$$a^- - a = \frac{\sum_{i<j}(X_i^c - X_j^c)(\lambda_i - \lambda_j) - \sum_{i<j}(X_i^r - X_j^r)(\eta_i - \eta_j)}{S^2(X^c)+S^2(X^r)}.$$

These imply

$$E(a^+ - a) = -\frac{2aS^2(X^r)}{S^2(X^c)+S^2(X^r)},$$
$$E(a^- - a) = 0.$$

Similar to the case of $a \geq 0$, we obtain

$$E(\hat{a}-a) = E(a^+ - a^-)I_{\{\hat{a}=a^+\}},$$
$$E(|\hat{a}|-|a|) = E(a^+ + a^-)I_{\{\hat{a}=a^+\}}.$$

These, together with (39) and (41), imply,

$$E(\hat{a}-a) = -\frac{2aS^2(X^r)}{S^2(X^c)+S^2(X^r)}P(\hat{a}=a^+), \tag{43}$$

$$E(|\hat{a}|-|a|) = \frac{2aS^2(X^c)}{S^2(X^c)+S^2(X^r)}P(\hat{a}=a^+). \tag{44}$$

The desired result follows from (40), (42), (43) and (44). □

5.5 Proof of Theorem 4

Proof. From (22) and (25),

$$E(\hat{b}|X) = E(\overline{Y^c} - \hat{a}\overline{X^c}|X) = E(a\overline{X^c} + b + \overline{\lambda} - \hat{a}\overline{X^c}|X) = \overline{X^c}E(a-\hat{a}|X) + b.$$

Similarly, from (23) and (26),

$$E(\hat{\mu}|X) = E(\overline{Y^r} - |\hat{a}|\overline{X^r}|X) = E(|a|\overline{X^r} + \overline{\eta} + \overline{\lambda} - |\hat{a}|\overline{X^r}|X) = \overline{X^r}E(|a|-|\hat{a}|\big|X) + \mu.$$

Hence, the desired result follows by Proposition 2 and Lemma 3 in the Appendix. □

6. Appendix II: Lemmas

Lemma 1. *Assume model (6)-(7) and* $Var(X^r) < \infty$. *Then* $Cov(X^r, Y^r) \geq 0$. *Consequently,* $S(X^r, Y^r) \geq 0$ *with probability converging to 1.*

Proof. According to (7),

$$\begin{aligned}
\mathrm{Cov}(X^r, Y^r) &= E(X^r Y^r) - E(X^r)E(Y^r) \\
&= E[X^r(|a|X^r + \eta_1)] - E(X^r)E(|a|X^r + \eta_1) \\
&= |a|E(X^r)^2 + \mu E(X^r) - |a|[E(X^r)]^2 - \mu E(X^r) \\
&= |a|\mathrm{Var}(X^r) \\
&\geq 0,
\end{aligned} \tag{45}$$

provided that $\mathrm{Var}(X^r) < \infty$. By the SLLN,

$$S(X^r, Y^r) \to \mathrm{Cov}(X^r, Y^r) \ a.s.. \tag{46}$$

(46) together with (45) completes the proof. □

Lemma 2. *The following are true for $v \in \{c, r\}$:*

$$S(X^v, Y^v) = \frac{1}{n^2} \sum_{i<j} (X_i^v - X_j^v)(Y_i^v - Y_j^v), \tag{47}$$

$$S^2(X^v) = \frac{1}{n^2} \sum_{i<j} (X_i^v - X_j^v)^2. \tag{48}$$

Proof. To prove (47),

$$\begin{aligned}
&\sum_{i<j} (X_i^v - X_j^v)(Y_i^v - Y_j^v) = \sum_{i<j} (X_i^v Y_i^v - X_i^v Y_j^v - X_j^v Y_i^v + X_j^v Y_j^v) \\
&= \sum_{i<j} (X_i^v Y_i^v + X_j^v Y_j^v) - \sum_{i<j} (X_i^v Y_j^v + X_j^v Y_i^v) \\
&= (n-1) \sum_{i=1}^n X_i^v Y_i^v - [(\sum_{i=1}^n X_i^v)(\sum_{i=1}^n Y_i^v) - \sum_{i=1}^n X_i^v Y_i^v] \\
&= n \sum_{i=1}^n X_i^v Y_i^v - (\sum_{i=1}^n X_i^v)(\sum_{i=1}^n Y_i^v) = n^2 S(X^v, Y^v).
\end{aligned}$$

(48) follows by replacing Y_i^v with X_i^v and Y_i^v with X_j^v in the above calculations. □

Lemma 3. *Assume model (6)-(7). Assume in addition that $S^2(X^c) = O(1)$ and $S^2(X^r) = O(1)$. Let $\{\hat{a}, \hat{b}, \hat{\mu}\}$ be the least squares solution defined in (13). Then*

$$\begin{aligned}
P(\hat{a} = a^- | a \geq 0) &\to 0, \\
P(\hat{a} = a^+ | a < 0) &\to 0,
\end{aligned}$$

as $n \to \infty$.

Proof. We prove the case $a \geq 0$ only. The case $a < 0$ can be proved similarly. Under the assumption that $a \geq 0$,

$$\mathrm{Cov}\left(X^c, Y^c\right) = a\mathrm{Var}\left(X^c\right) \geq 0,$$

and consequently, $P\left(S\left(X^c, Y^c\right) < 0\right) \to 0$. According to Theorem 3, the only other circumstance under which $\hat{a} = a^-$ is when $S\left(X^r, Y^r\right) > S\left(X^c, Y^c\right) > 0$ and $L\left(a^+, b^+, \mu^+\right) > L\left(a^-, b^-, \mu^-\right)$ simultaneously. It is therefore sufficient to show that

$$P\left(S\left(X^r, Y^r\right) > S\left(X^c, Y^c\right) > 0, L\left(a^+, b^+, \mu^+\right) > L\left(a^-, b^-, \mu^-\right)\right) \to 0. \quad (49)$$

Notice

$$\begin{aligned} & L\left(a^+, b^+, \mu^+\right) - L\left(a^-, b^-, \mu^-\right) \\ = \quad & \frac{1}{n}\sum_{i=1}^{n}\left[\left(a^+X_i^c + b - Y_i^c\right)^2 - \left(a^-X_i^c + b - Y_i^c\right)^2\right] \\ & + \frac{1}{n}\sum_{i=1}^{n}\left[\left(a^+X_i^r + \mu - Y_i^r\right)^2 - \left(a^-X_i^r + \mu - Y_i^r\right)^2\right] \\ := \quad & \frac{1}{n}\left(I + II\right). \end{aligned}$$

The first term

$$\begin{aligned} I &= \sum_{i=1}^{n}\left[\left(a^+X_i^c + b - Y_i^c\right)^2 - \left(a^-X_i^c + b - Y_i^c\right)^2\right] \\ &= \sum_{i=1}^{n}\left[\left(a^+ - a\right)^2\left(X_i^c - \overline{X^c}\right)^2 + \left(\lambda_i - \overline{\lambda}\right)^2 - 2\left(a^+ - a\right)\left(X_i^c - \overline{X^c}\right)\left(\lambda_i - \overline{\lambda}\right)\right] \\ &\quad - \sum_{i=1}^{n}\left[\left(a^- - a\right)^2\left(X_i^c - \overline{X^c}\right)^2 + \left(\lambda_i - \overline{\lambda}\right)^2 - 2\left(a^- - a\right)\left(X_i^c - \overline{X^c}\right)\left(\lambda_i - \overline{\lambda}\right)\right] \\ &= \left[\left(a^+ - a\right)^2 - \left(a^- - a\right)^2\right]\sum_{i=1}^{n}\left(X_i^c - \overline{X^c}\right)^2 \\ &\quad - 2\left(a^+ - a^-\right)\sum_{i=1}^{n}\left(X_i^c - \overline{X^c}\right)\left(\lambda_i - \overline{\lambda}\right) \\ &= \left(a^+ - a^-\right)\left[\left(a^+ + a^- - 2a\right)\sum_{i=1}^{n}\left(X_i^c - \overline{X^c}\right)^2 - 2\sum_{i=1}^{n}\left(X_i^c - \overline{X^c}\right)\left(\lambda_i - \overline{\lambda}\right)\right]. \end{aligned}$$

From this, and the assumption that $S(X^r,Y^r)>S(X^c,Y^c)>0$, we see that $I>0$ is equivalent to

$$\left(\frac{a^+ + a^-}{2} - a\right)\sum_{i=1}^{n}\left(X_i^c - \overline{X^c}\right)^2 - \sum_{i=1}^{n}\left(X_i^c - \overline{X^c}\right)\left(\lambda_i - \overline{\lambda}\right) > 0. \tag{50}$$

On the other hand,

$$\begin{aligned}
&(50)\\
&= \left[\frac{S(X^c,Y^c)}{S^2(X^c)+S^2(X^r)} - a\right]\sum_{i=1}^{n}\left(X_i^c - \overline{X^c}\right)^2 - \sum_{i=1}^{n}\left(X_i^c - \overline{X^c}\right)\left(\lambda_i - \overline{\lambda}\right)\\
&= \left[\frac{\sum_{i<j}\left(X_i^c - X_j^c\right)(\lambda_i - \lambda_j)}{\sum_{i<j}\left(X_i^c - X_j^c\right)^2 + \sum_{i<j}\left(X_i^r - X_j^r\right)^2} - a\frac{S^2(X^r)}{S^2(X^c)+S^2(X^r)}\right]\sum_{i=1}^{n}\left(X_i^c - \overline{X^c}\right)^2\\
&- \sum_{i=1}^{n}\left(X_i^c - \overline{X^c}\right)\left(\lambda_i - \overline{\lambda}\right)\\
&= \frac{\sum_{i=1}^{n}\left(X_i^c - \overline{X^c}\right)^2}{\sum_{i<j}\left(X_i^c - X_j^c\right)^2 + \sum_{i<j}\left(X_i^r - X_j^r\right)^2}\sum_{i<j}\left(X_i^c - X_j^c\right)(\lambda_i - \lambda_j)\\
&- \sum_{i=1}^{n}\left(X_i^c - \overline{X^c}\right)\left(\lambda_i - \overline{\lambda}\right) - a\frac{S^2(X^r)}{S^2(X^c)+S^2(X^r)}\sum_{i=1}^{n}\left(X_i^c - \overline{X^c}\right)^2\\
&= \frac{\sum_{i=1}^{n}\left(X_i^c - \overline{X^c}\right)^2}{\sum_{i<j}\left(X_i^c - X_j^c\right)^2 + \sum_{i<j}\left(X_i^r - X_j^r\right)^2}\left[n\sum_{i=1}^{n}\left(X_i^c - \overline{X^c}\right)\left(\lambda_i - \overline{\lambda}\right)\right]\\
&- \sum_{i=1}^{n}\left(X_i^c - \overline{X^c}\right)\left(\lambda_i - \overline{\lambda}\right) - a\frac{S^2(X^r)}{S^2(X^c)+S^2(X^r)}\sum_{i=1}^{n}\left(X_i^c - \overline{X^c}\right)^2\\
&= \sum_{i=1}^{n}\left(X_i^c - \overline{X^c}\right)\left(\lambda_i - \overline{\lambda}\right)\left[\frac{S^2(X^c)}{S^2(X^c)+S^2(X^r)} - 1\right]\\
&- a\frac{S^2(X^r)}{S^2(X^c)+S^2(X^r)}\sum_{i=1}^{n}\left(X_i^c - \overline{X^c}\right)^2\\
&= -\frac{S^2(X^r)}{S^2(X^c)+S^2(X^r)}n\left[aS^2(X^c) + S(X^c,\lambda)\right],
\end{aligned}$$

where $S(X^c,\lambda)=\frac{1}{n}\sum_{i=1}^{n}\left(X_i^c-\overline{X^c}\right)\left(\lambda_i-\overline{\lambda}\right)$ denotes the sample covariance of the random variables X^c and λ, which converges to 0 almost surely by the independence assumption. Therefore,

$$\begin{aligned}\frac{1}{n}I &= -2\left(a^+-a^-\right)\frac{S^2(X^r)}{S^2(X^c)+S^2(X^r)}\left[aS^2(X^c)+S(X^c,\lambda)\right]\\ &\to C_1<0\end{aligned} \tag{51}$$

almost surely, as $n\to\infty$.

By the similar calculation, we have that the second term

$$\begin{aligned}\frac{1}{n}II &= -2\left(|a^+|-|a^-|\right)\frac{S^2(X^c)}{S^2(X^c)+S^2(X^r)}\left[aS^2(X^r)+S(X^r,\eta)\right]\\ &\to C_2<0\end{aligned} \tag{52}$$

almost surely, as $n\to\infty$. (51) and (52) together imply that

$$P\left(\hat{a}=a^-|a\geq 0\right)\to 0.$$

This completes the proof. □

References

[1] Artstein, Z, & Vitale, R.A. (1975). A strong law of large numbers for random compact sets. *Annals of Probability, 5*, 879-882.

[2] Aumann, R.J. (1965). Integrals of set-valued functions. *J. Math. Anal. Appl., 12*,1-12.

[3] Billard, L., & Diday, E. (2000). Regression analysis for interval-valued data. In: *Data Analysis, Classification and Related Methods, Proceedings of the Seventh Conference of the International Federation of Classification Societies (IFCS'00)*. Springer, Berlin; 369-374.

[4] Billard, L., & Diday, E. (2002). Symbolic regression analysis. In: *Classification, Clustering and Data Analysis, Proceedings of the Eighth Conference of the International Federation of Classification Societies (IFCS'02)*. Springer, Berlin; 281-288.

[5] Billard, L. (2007). Dependencies and variation components of symbolic interval-valued data. In: *Selected Contributions in Data Analysis and Classification*. Springer, Berlin Heidelberg; 3-12.

[6] Blanco-Fernández, A., Corral, N., & González-Rodríguez, G. (2011). Estimation of a flexible simple linear model for interval data based on set arithmetic. *Computational Statistics & Data Analysis, 55*, 2568-2578.

[7] Blanco-Fernández, A., Colubi, A., & González-Rodríguez, G. (2012). Confidence sets in a linear regression model for interval data. *Journal of Statistical Planning and Inference, 142*, 1320-1329.

[8] Carvalho, F.A.T., Lima Neto, E.A., & Tenorio, C.P. (2004). A new method to fit a linear regression model for interval-valued data. *Lecture Notes in Computer Sciences, 3238*, 295-306.

[9] Cattaneo, M.E.G.V., & Wiencierz, A. (2012). Likelihood-based imprecise regression. *International Journal of Approximate Reasoning, 53*, 1137-1154.

[10] Diamond, P. (1990). Least squares fitting of compact set-valued data. *J. Math. Anal. Appl., 147*, 531-544.

[11] Gil, M.A., Lopez, M.T., Lubiano, M.A., & Montenegro, M. (2001). Regression and correlation analyses of a linear relation between random intervals. *Test,10*, 183-201.

[12] Gil, M.A., Lubiano, M.A., Montenegro, M., & Lopez, M.T. (2002). Least squares fitting of an affine function and strength of association for interval-valued data. *Metrika, 56*, 97-111.

[13] Gil, M.A., González-Rodríguez, G., Colubi, A., & Montenegro, M. (2007). Testing linear independence in linear models with interval-valued data. *Computational Statistics & Data Analysis, 51*, 3002-3015.

[14] González-Rodríguez, G., Blanco, A., Corral, N., & Colubi, A. (2007). Least squares estimation of linear regression models for convex compact random sets. *Advances in Data Analysis and Classification, 1*, 67-81.

[15] Hörmander, H. (1954). Sur la fonction d'appui des ensembles convexes dans un espace localement convexe. *Arkiv för Mat, 3*, 181-186.

[16] Hukuhara, M. (1967). Integration des applications mesurables dont la valeur est un compact convexe. *Funkcialaj Ekvacioj, 10*, 205-223.

[17] Kendall, D.G. (1974). Foundations of a theory of random sets. In: Harding EF, & Kendall DG (Eds), *Stochastic Geometry*. New York: John Wiley & Sons.

[18] Körner, R. (1995). A variance of compact convex random sets. Institut für Stochastik, Bernhard-von-Cotta-Str. 2 09599 Freiberg.

[19] Körner, R. (1997). On the variance of fuzzy random variables. *Fuzzy Sets and Systems, 92*, 83-93.

[20] Körner, R., & Näther, W. (1998). Linear regression with random fuzzy variables: extended classical estimates, best linear estimates, least squares estimates. *Information Sciences, 109*, 95-118.

[21] Lyashenko, N.N. (1982). Limit theorem for sums of independent compact random subsets of Euclidean space. *Journal of Soviet Mathematics, 20*, 2187-2196.

[22] Lyashenko, N.N. (1983). Statistics of random compacts in Euclidean space. *Journal of Soviet Mathematics, 21*, 76-92.

[23] Manski, C.F., & Tamer, T. (2002). Inference on regressions with interval data on a regressor or outcome. *Econometrica, 70*, 519-546.

[24] Matheron, G. (1975). *Random Sets and Integral Geometry*. New York: John Wiley & Sons.

[25] Molchanov, I. (2005). *Theory of Random Sets*. London: Springer.

[26] Lima Neto, E.A., & Carvalho, F.A.T. (2008). Centre and range method for fitting a linear regression model to symbolic interval data. *Computational Statistics & Data Analysis, 52*, 1500-1515.

[27] Lima Neto, E.A., & Carvalho, F.A.T. (2010). Constrained linear regression models for symbolic interval-valued variables. *Computational Statistics & Data Analysis, 54*,333-347.

[28] Rådström, H. (1952) An embedding theorem for spaces of convex sets. *Proc. Amer. Math. Soc., 3*, 165-169.

In: Linear Regression
Editor: Vera L. Beck
ISBN: 978-1-53611-992-3
© 2017 Nova Science Publishers, Inc.

Chapter 4

LINEAR REGRESSION VERSUS NON-LINEAR REGRESSION IN MATHEMATICAL MODELING OF ADSORPTION PROCESSES

Gabriela-Nicoleta Moroi*, PhD
Laboratory of Polyaddition and Photochemistry
"Petru Poni" Institute of Macromolecular Chemistry
Iaşi, Romania

ABSTRACT

In mathematical modeling of adsorption processes, linear and/or non-linear regression analysis may be employed. In adsorption isotherm modeling, non-linear regression has lately been reported by some authors to provide a better fit to experimental data than linear regression. Isotherm models used in describing the adsorption systems, criteria selected to evaluate isotherm model validity as well as modeling results are comparatively discussed.

In our investigation on modeling of adsorption of heavy metal ions onto surface-functionalized polymer beads, linear and non-linear

* Corresponding Author: gmoroi@icmpp.ro.

regressions were employed for each of the isotherm models considered to describe the equilibrium data. To reliably assess model validity, various error functions (whose mathematical expressions contain the number of experimental measurements, the numbers of independent variables and parameters in the regression equation as well as the measured and predicted equilibrium adsorption capacities) were used. The modeling results obtained by employing the two regression methods were compared. For the adsorption of each metal ion species, it was revealed that (a) for a particular isotherm model, the regression providing the best fit is linear, non-linear or both linear and non-linear, and (b) the order of isotherm model validities indicated via linear regression is the same with that shown by non-linear regression.

Keywords: adsorption isotherm modeling, linear regression, non-linear regression, heavy metal ions, surface-functionalized polymer beads, ionic liquid-like functionalities

INTRODUCTION

Heavy metals are highly toxic environmental pollutants making a great impact on human health and environment quality, which are usually introduced into natural water resources by wastewaters resulting from industrial activities; therefore, removal of heavy metals from contaminated waters is an *absolute necessity for* public health protection and environmental conservation (Sigel et al. 2013; Casas and Sordo 2006). Adsorption is one of the most popular procedures used in wastewater treatment for preventing environmental contamination. Materials with adsorption ability towards metals can be obtained by chemically immobilizing functional groups onto polymeric supports; e.g., styrene-divinylbenzene copolymer beads with ionic liquid-like functionalities (1-methyl-3-methylimidazolium chloride) covalently attached onto their surface (ILLF-SDVB) were synthesized and employed to remove *heavy metal ions* from aqueous solutions (Moroi et al. 2016; Moroi 2012; Bilba et al. 2007; Moroi et al. 2006; Bilba et al. 2006; Moroi et al. 2004; Bilba et al. 2004; Moroi et al. 2001).

Generally, the equilibrium state of an adsorption process is characterized by the relationship between the amount of adsorbate being adsorbed and the amount of adsorbate remaining in solution. An experimental equilibrium isotherm, i.e., equilibrium adsorption capacity (q_e) versus equilibrium adsorbate concentration in solution (C_e) plot, reflects the change of adsorbate distribution between adsorbent and solution as C_e increases, at constant temperature and pH; such an isotherm may be analyzed by various isotherm models for determining which model provides the best mathematical description of experimental data and the best prediction of adsorption parameters. Finding the best-fitting model is of great importance since the thermodynamic assumptions and parameter estimates give information on adsorbent surface properties, adsorbent-adsorbate affinity and adsorption mechanism that are useful for optimizing adsorption system design. In mathematical modeling of equilibrium adsorption isotherms, linear and/or non-linear regression analysis may be employed. A large variety of modeling approaches are used that differ from each other as regards the number and type of (a) isotherm models considered (Langmuir, Freundlich, Dubinin–Radushkevich, Temkin, Flory–Huggins, Hill, Redlich–Peterson, Sips, Koble–Corrigan, Toth etc.), (b) error functions minimized/maximized, (c) error functions calculated and (d) criteria based on the calculated error functions that are employed to assess isotherm model validity (Foo and Hameed 2010, Han et al. 2009, Ho et al. 2002).

COMMENTS ON ARTICLES STATING THAT NON-LINEAR REGRESSION IS BETTER THAN LINEAR REGRESSION IN ADSORPTION ISOTHERM MODELING

The following observations were made on several articles that compare the results of linear regression and non-linear regression in modeling of adsorption isotherms. The names of isotherm parameters and error functions are used inconsistently, which may cause confusion (Armagan

and Toprak 2013, Kumar 2006). Some mathematical expressions of isotherm models and error functions are written incorrectly or not shown at all, making questionable the accuracy of modeling results (Brdar et al. 2012, Kumar et al. 2008). The main observation regards the statement that non-linear regression provides better results than linear regression, which however is in disagreement with the presented data; some examples of such discrepancies are given below using the names and abbreviations employed in the articles.

In the study of Cu(II) adsorption onto lignin by linear and non-linear regression analysis of Freundlich, Langmuir and Redlich–Peterson isotherm models, two different modeling approaches are employed for each isotherm model: in linear regression, least square method is used to calculate one value of r^2 and one value of chi-square test, whereas in non-linear regression, the values of five error functions, i.e., ERRSQ, HYBRD, MPSE, ARE and EABS, are minimized to obtain five values of r^2 and five values of chi-square test (Brdar et al. 2012). It is noted that the statement that non-linear regression is better than linear regression is not supported by r^2 and chi-square test values, these indicating, on the contrary, that linear regression is comparatively better; e.g., in the case of Redlich–Peterson isotherm model, on one hand, r^2 value for linear regression is higher than the following r^2 values for non-linear regression: each value corresponding to HYBRD, MPSE, ARE and EABS, the average value of the three higher values that correspond to ERRSQ, HYBRD and MPSD and the average value of the five values corresponding to each error function and, on the other hand, chi-square test value for linear regression is lower than the following chi-square test values for non-linear regression: each value corresponding to ERRSQ, MPSE, ARE and EABS, the average value of the three lower values that correspond to ERRSQ, HYBRD and MPSD and the average value of the five values corresponding to each error function. The very good fit to experimental data of linearized Redlich–Peterson isotherm model is also graphically revealed, whereas such a figure is not shown for non-linearized Redlich–Peterson isotherm model.

In the comparative investigation of linear and non-linear regressions to estimate isotherm parameters for adsorption of malachite green onto

activated carbon, it is stated that linear method is inappropriate for describing adsorption isotherm and it is better to use non-linear method ("which have a uniform error distribution (irrespective of the linear form) for the whole range of experimental data") (Kumar 2006). However, of all 24 values of r^2 calculated by using linear regression (the experimental data obtained at four temperatures being analyzed by Langmuir in four forms, Freundlich and Redlich-Peterson isotherm models), 15 values, representing more than half of the total number of values, are higher than the corresponding values calculated by employing non-linear regression.

In studying the adsorption of methylene blue onto activated carbon by Langmuir, Freundlich and Redlich–Petersen isotherm models, different approaches are employed for linear regression (r^2 and least squares method) and non-linear regression (six error functions, i.e., r^2, ERRSQ, HYBRID, MPSD, ARE and EABS, and a trial and error method) (Kumar 2008). It is stated that non-linear regression is a better way compared with linear regression to obtain isotherm parameters and select the optimum isotherm ("as sometime linearization of non-linear experimental data may distort the error distribution structure of isotherm"). However, the same conclusion is reached by the two regression methods, i.e., that this adsorption process is "well represented by both Langmuir and Redlich Peterson isotherm."

In isotherm modeling of NaCN adsorption onto activated carbon by using six isotherm models (Langmuir, Freundlich, Dubinin–Radushkevich, Temkin, Redlich–Peterson and Koble–Corrigan) and three error functions (R^2, MPSD and HYBRID), it is stated that non-linear regression is better than linear regression for predicting isotherm parameters (Salarirad and Behnamfard 2011). However, of 18 pairs of values of all error functions obtained by using linear and non-linear regressions for all isotherm models considered, in only 8 pairs, i.e., in less than half of the total number of pairs, the value provided by non-linear regression is better than that given by linear regression, whereas in 9 pairs, on the contrary, linear regression value is better than non-linear regression value and, in one pair, the linear and non-linear regression values are equal.

The above comments highlight the need for reliable, consistent approaches to assess the performance of linear and non-linear regressions in accurately describing adsorption processes by using isotherm models.

MODELING BY LINEAR AND NON-LINEAR REGRESSIONS OF EQUILIBRIUM ISOTHERMS IN ADSORPTION OF HEAVY METAL IONS ONTO SURFACE-FUNCTIONALIZED POLYMER BEADS

In performing Cd(II) and Pb(II) (Me) adsorption onto ILLF-SDVB beads, aqueous solutions of the toxic cadmium nitrate, $Cd(NO_3)_2{\cdot}4H_2O$, and lead nitrate, $Pb(NO_3)_2$ were used. Batch experiments were carried out at 20ºC, pH of 5, adsorbent dose of 4.100 g L^{-1}, contact time of 24 h and different initial Me concentrations in solution (C_0) varying in the ranges of 0.200–2.810 and 0.106–1.700 mmol L^{-1} for Cd(II) and Pb(II), respectively; equilibrium Me concentrations in solution (C_e) were spectrophotometrically determined (Moroi et al. 2016). Adsorption performance was evaluated by equilibrium adsorption capacity (q_e):

$$q_e = \frac{(C_0 - C_e)V}{m} \text{ (mmol g}^{-1}\text{)} \quad (1)$$

where C_0 and C_e – the initial and equilibrium adsorbate (Me) concentrations in solution, respectively (mmol L^{-1}); V – the volume of solution (L); m – the mass of adsorbent (g).

As already shown, in experimental equilibrium isotherm modeling, non-linear regression has been reported by some authors to provide a better fit to experimental data than linear regression; to determine whether this statement is valid for adsorption of Cd(II) and Pb(II) onto ILLF-SDVB beads, both linear and non-linear regressions were employed in the present study to analyze the equilibrium data by the two-parameter Langmuir,

Temkin and Freundlich isotherm models. For consistency reasons, the same approach was used for both regressions: the same nine error functions were calculated by minimizing the sum of squared *errors* (SSE) and the same two criteria based on these error functions were employed to establish isotherm model validities. The main features of the three isotherm models used in mathematical modeling of Cd(II) and Pb(II) adsorption are presented below.

Langmuir isotherm model is based on the assumption of a homogeneous adsorption with monomolecular layer coverage of a surface with a finite number of energetically equivalent sites, one site being occupied by only one adsorbate species; there is no interaction among adsorbed species and no transmigration of adsorbed species in the surface plane (Langmuir 1916). Four linear forms, whose plots are C_e/q_e versus C_e, $1/q_e$ versus $1/C_e$, q_e versus q_e/C_e and q_e/C_e versus q_e (Llin1, Llin2, Llin3 and Llin4, respectively), and non-linear form (Lnonlin) of Langmuir isotherm are displayed below:

- Llin1:

$$\frac{C_e}{q_e} = \frac{1}{q_m} C_e + \frac{1}{K_L q_m} \tag{2}$$

- Llin2:

$$\frac{1}{q_e} = \frac{1}{K_L q_m} \frac{1}{C_e} + \frac{1}{q_m} \tag{3}$$

- Llin3:

$$q_e = -\frac{1}{K_L} \frac{q_e}{C_e} + q_m \tag{4}$$

- Llin4:

$$\frac{q_e}{C_e} = -K_L\, q_e + K_L\, q_m \tag{5}$$

- Lnonlin:

$$q_e = \frac{q_m K_L C_e}{1 + K_L C_e} \tag{6}$$

where q_e – the measured equilibrium adsorption capacity (mmol g^{-1}); C_e – the measured equilibrium adsorbate concentration in solution (mmol L^{-1}); q_m – the Langmuir isotherm constant representing the maximum adsorption capacity (complete monolayer coverage) (mmol g^{-1}) and K_L – the Langmuir isotherm constant (adsorbent-adsorbate affinity parameter) related to binding energy (L $mmol^{-1}$).

By employing Langmuir constant K_L, two important adsorption parameters are assessed:

- the dimensionless separation factor (R_L), whose value may be 0, between 0 and 1, 1 or above 1, indicating whether adsorption nature is irreversible, favorable, linear or unfavorable, respectively (McKay et al. 1982, Wasewar 2010):

$$R_L = \frac{1}{1 + K_L C_{0h}} \tag{7}$$

where K_L – the Langmuir isotherm constant (L $mmol^{-1}$) and C_{0h} – the highest initial concentration of adsorbate in solution (mmol L^{-1})

- the *Gibbs free* energy change (ΔG^0) indicates adsorption feasibility and spontaneous nature (He et al. 2010):

$$\Delta G^0 = -R\,T\ln K_L \ (\text{J mol}^{-1}) \tag{8}$$

where K_L – the Langmuir isotherm constant (L mol^{-1}); R – the universal gas constant (8.314 J mol^{-1} K^{-1}) and T – the absolute temperature (K).

Temkin isotherm model assumes that adsorption heat of molecules decreases linearly with increasing surface coverage due to adsorbent-adsorbate interactions, adsorption being characterized by a uniform distribution of binding energies up to a maximum energy value; a good fit of Temkin isotherm to experimental equilibrium data reveals the occurrence of chemisorption (Temkin 1941; Foo and Hameed 2010; Boparai et al. 2011). Linear form, whose plot is q_e versus $\ln(C_e)$ (Tlin), and non-linear form (Tnonlin) of Temkin isotherm are shown nextly:

- Tlin:

$$q_e = \frac{R\,T}{b_T}\ln(C_e) + \frac{R\,T}{b_T}\ln(K_T) \tag{9}$$

- Tnonlin:

$$q_e = \frac{R\,T}{b_T}\ln(K_T\,C_e) \tag{10}$$

where q_e – the measured equilibrium adsorption capacity (mmol g^{-1}); C_e – the measured equilibrium adsorbate concentration in solution (mmol L^{-1}); b_T – the Temkin isotherm constant related to adsorption heat (kJ mol^{-1}); K_T – the Temkin equilibrium binding constant corresponding to the maximum binding energy (L $mmol^{-1}$); R – the universal gas constant (8.314 J mol^{-1} K^{-1}) and T – the absolute temperature (K).

Freundlich isotherm model hypothesizes a multiple layer adsorption on an energetically heterogeneous surface and a logarithmic decrease in adsorption energy with increasing surface coverage (Freundlich 1906; Ho

and McKay 1998). Linear form, whose plot is $\ln(q_e)$ versus $\ln(C_e)$ (Flin), and non-linear form (Fnonlin) of Freundlich isotherm are presented below:

- Flin:

$$\ln(q_e) = \frac{1}{n}\ln(C_e) + \ln(K_F) \tag{11}$$

- Fnonlin:

$$q_e = K_F C_e^{1/n} \tag{12}$$

where q_e – the measured equilibrium adsorption capacity (mmol g^{-1}); C_e – the measured equilibrium adsorbate concentration in solution (mmol L^{-1}); K_F – the Freundlich isotherm constant indicative of adsorption capacity ($mmol^{1-1/n}$ $L^{1/n}$ g^{-1}) and 1/n – Freundlich isotherm constant related to adsorption intensity, representing a measure of surface energetic heterogeneity.

In mathematical modeling of experimental equilibrium isotherm data of Cd(II) and Pb(II) adsorption onto ILLF-SDVB beads, parameter estimates for Langmuir, Temkin and Freundlich models were calculated by using linear and non-linear least-squares regression analysis, minimizing SSE:

$$SSE = \sum_{i=1}^{n} (q_e - \hat{q}_e)_i^2 \tag{13}$$

where n – the number of experimental measurements and q_e and $\hat{q}_e$ – the measured and predicted equilibrium adsorption capacities, respectively.

For reliably assessing the validity of each linear and non-linear isotherm model form for the studied adsorption systems (i.e., the goodness of fit between $\hat{q}_e$ and q_e), nine error functions, which are either relative or absolute, were employed (Ho et al. 2002; Foo and Hameed 2010). The mathematical expressions of these error functions are presented next.

- average relative error (ARE):

$$\mathrm{ARE} = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{q_e - \hat{q}_e}{q_e}\right|_i \tag{14}$$

- average absolute error (EABS):

$$\mathrm{EABS} = \frac{1}{n}\sum_{i=1}^{n}\left|q_e - \hat{q}_e\right|_i \tag{15}$$

- sum of the absolute errors (SAE):

$$\mathrm{SAE} = \sum_{i=1}^{n}\left|q_e - \hat{q}_e\right|_i \tag{16}$$

- root mean square error (RMSE):

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(q_e - \hat{q}_e\right)_i^2} \tag{17}$$

- Marquardt's percent standard deviation (MPSD):

$$\mathrm{MPSD} = 100\sqrt{\frac{1}{n-p}\sum_{i=1}^{n}\left(\frac{q_e - \hat{q}_e}{q_e}\right)_i^2} \tag{18}$$

- average relative standard error (ARSE):

$$\mathrm{ARSE} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{q_e - \hat{q}_e}{q_e}\right)_i^2} \tag{19}$$

- hybrid fractional error function (HYBRID):

$$\mathrm{HYBRID} = \frac{100}{n-p}\sum_{i=1}^{n}\left[\frac{(q_e - \hat{q}_e)^2}{q_e}\right]_i \tag{20}$$

- chi-square test (CST):

$$\mathrm{CST} = \sum_{i=1}^{n}\left[\frac{(q_e - \hat{q}_e)^2}{\hat{q}_e}\right]_i \tag{21}$$

- adjusted coefficient of determination (ADRSQ):

$$\mathrm{ADRSQ} = 1 - (1 - R^2)\left[\frac{n-1}{n-(k+1)}\right] \tag{22}$$

where n – the number of experimental measurements; k and p – the numbers of independent variables and parameters, respectively, in the regression equation; q_e, $\hat{q}_e$ and $\bar{q}_e$ – the measured, predicted and average measured equilibrium adsorption capacities, respectively, and R^2 – the coefficient of determination:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{q}_e - \bar{q}_e)_i^2}{\sum_{i=1}^{n}(q_e - \hat{q}_e)_i^2 + \sum_{i=1}^{n}(\hat{q}_e - \bar{q}_e)_i^2} \tag{23}$$

For all error functions except ADRSQ, the lower the value, the closer the match between $\hat{q}_e$ and q_e; for ADRSQ, whose values may vary from 0 to 1, a higher value indicates that $\hat{q}_e$ more closely match q_e.

After determining the values of all error functions for all linear and non-linear isotherm model forms, the following calculations were performed for comparison reasons:

- for every error function, percent deviation (EPD) of each of its values (E) with respect to the best of these values (E_0, which is the maximum value for ADRSQ and the minimum value for the other error functions) was determined:

$$EPD = 100\frac{|E - E_0|}{E_0} \ (\%) \tag{24}$$

- for each isotherm model form, the sum of EPD values of all error functions (SEPD) was calculated:

$$SEPD = \sum_{i=1}^{9} EPD \ (\%) \tag{25}$$

The validity of an isotherm model was estimated by employing two criteria that take into account all nine error functions, knowing that good criteria of validity are those based on a combination of relative and absolute error functions (Legates and McCabe 1999). The first criterion is the number of error functions having the minimum among EPD values of compared isotherm models (EF_{min}) and the second criterion is the SEPD value. Thus, the greater the number of EF_{min} and the smaller the value of SEPD, the better the model validity.

For Cd(II) and Pb(II) adsorption onto ILLF-SDVB beads, linear (Figures 1 and 2, respectively) and non-linear (Figures 3 and 4, respectively) forms of Langmuir, Temkin and Freundlich isotherm models

were considered (Moroi et al. 2016). Plots of experimental equilibrium data for both Me species reveal an increase in q_e with increasing C_e (Figures 3 and 4); these L-type isotherms indicate chemisorption, reflecting a high adsorbate-adsorbent affinity (Bradl 2004). Isotherm parameter estimates as well as R_L and ΔG^0 values were determined for adsorption of Cd(II) and Pb(II) (Tables 1 and 2, respectively). The values of error functions for Cd(II) and Pb(II) adsorption differ largely from each other, being spread over a wide numerical range (Tables 3 and 4, respectively); to reach reliable conclusions on the validity of isotherm models by analyzing comparable data, EPD and SEPD values were calculated (considering only the isotherm forms included in each table/figure) (Tables 5 and 6, respectively, and Figures 5 and 6, respectively).

Table 1. Values of isotherm model parameters for Cd(II) adsorption

Isotherm model	Adsorption parameters			
Langmuir	q_m (mmol g^{-1})	K_L (L $mmol^{-1}$)	R_L	ΔG^{0a} (kJ mol^{-1})
Llin1	0.112	3.60	0.581	–19.96
Llin2	0.116	3.42	0.594	–19.83
Llin3	0.109	4.04	0.553	–20.24
Llin4	0.112	3.68	0.576	–20.02
Lnonlin	0.111	3.87	0.564	–20.14
Temkin	b_T (kJ mol^{-1})	K_T (L $mmol^{-1}$)		
Tlin	110.03	45.17		
Tnonlin	110.02	45.16		
Freundlich	K_F ($mmol^{1-1/n}$ $L^{1/n}$ g^{-1})	1/n		
Flin	0.081	0.372		
Fnonlin	0.081	0.312		

[a] For C_{0h} = 2.810 mmol L^{-1}.

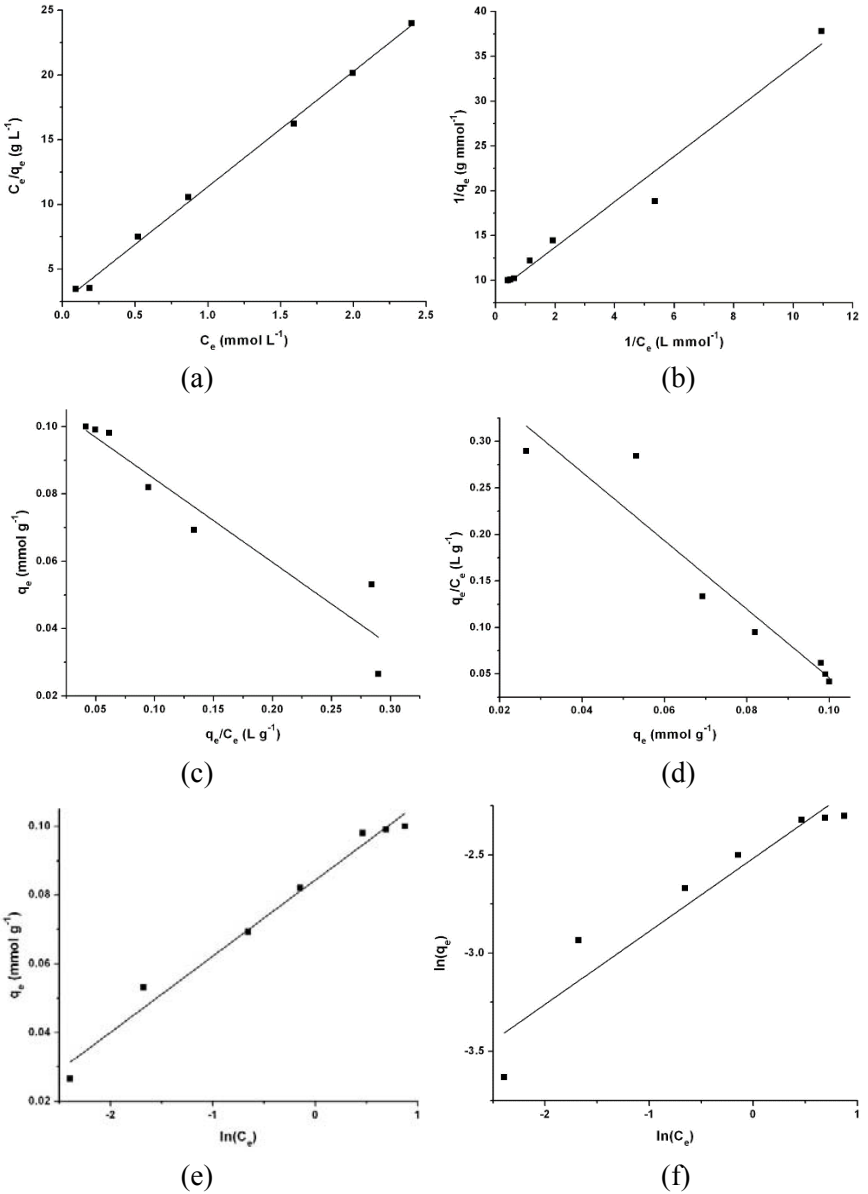

Figure 1. Linear Langmuir Llin1 (a), Langmuir Llin2 (b), Langmuir Llin3 (c), Langmuir Llin4 (d), Temkin (e) and Freundlich (f) isotherms for Cd(II) adsorption.

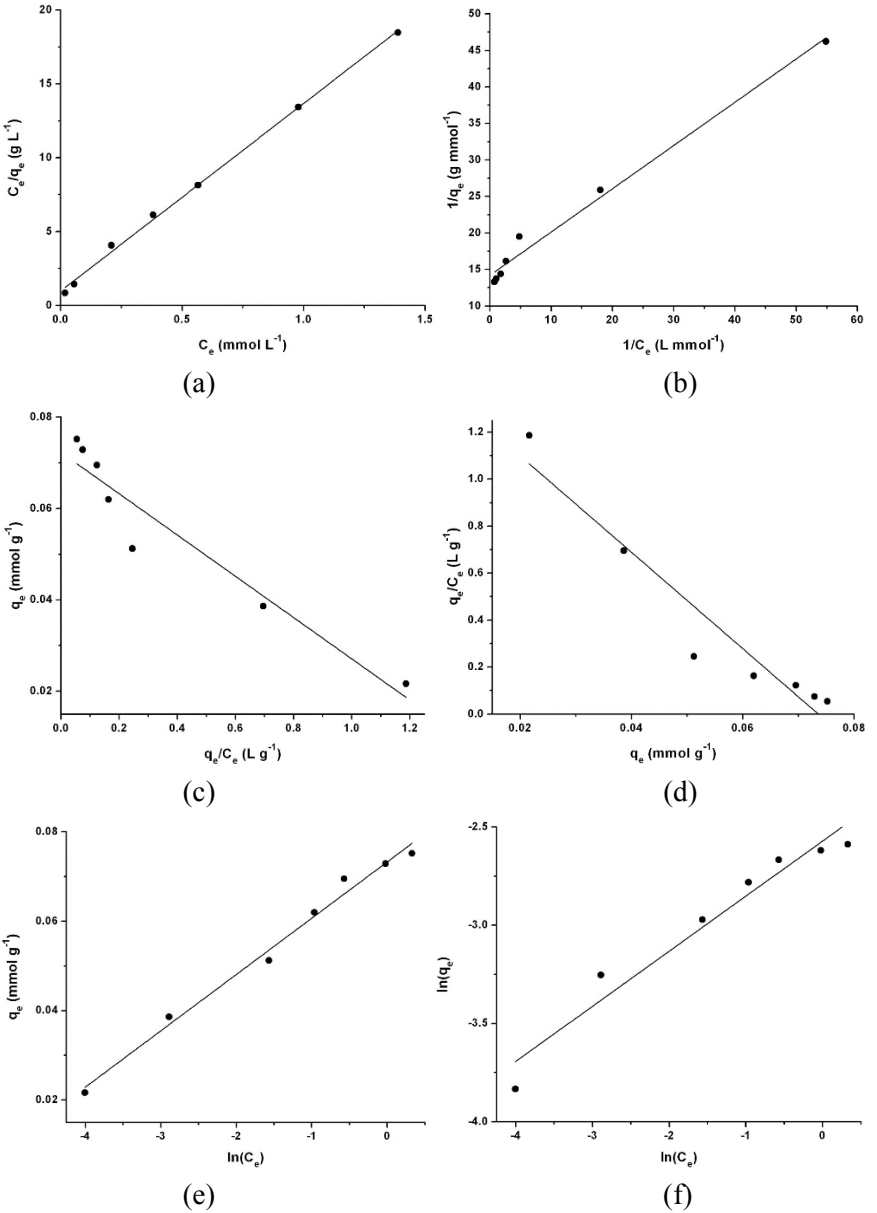

Figure 2. Linear Langmuir Llin1 (a), Langmuir Llin2 (b), Langmuir Llin3 (c), Langmuir Llin4 (d), Temkin (e) and Freundlich (f) isotherms for Pb(II) adsorption.

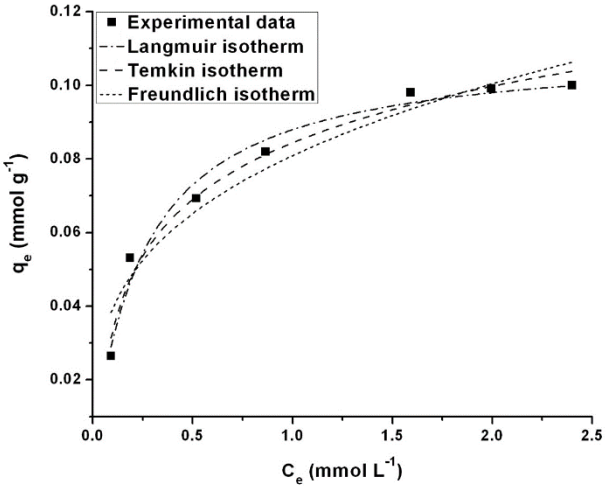

Figure 3. Experimental data and non-linear Langmuir, Temkin and Freundlich isotherms for Cd(II) adsorption.

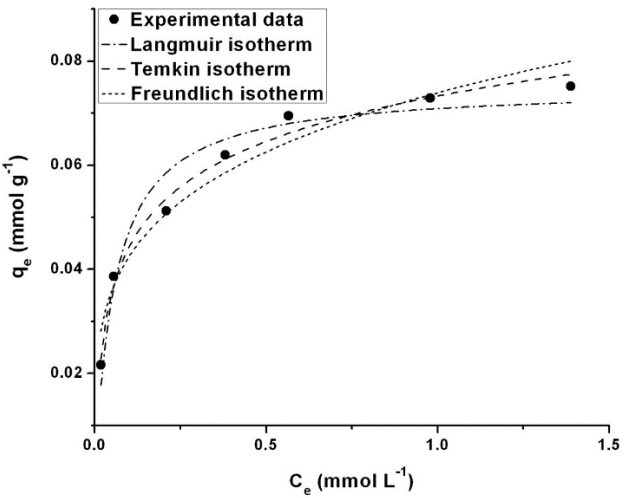

Figure 4. Experimental data and non-linear Langmuir, Temkin and Freundlich isotherms for Pb(II) adsorption.

Table 2. Values of isotherm model parameters for Pb(II) adsorption

Isotherm model	Adsorption parameters			
Langmuir	q_m (mmol g^{-1})	K_L (L $mmol^{-1}$)	R_L	ΔG^{0a} (kJ mol^{-1})
Llin1	0.079	12.96	0.421	–23.08
Llin2	0.070	24.00	0.282	–24.58
Llin3	0.072	22.13	0.299	–24.39
Llin4	0.074	20.52	0.315	–24.20
Lnonlin	0.075	17.00	0.357	–23.74
Temkin	b_T (kJ mol^{-1})	K_T (L $mmol^{-1}$)		
Tlin	193.43	336.12		
Tnonlin	193.45	336.30		
Freundlich	K_F ($mmol^{1-1/n}$ $L^{1/n}$ g^{-1})	1/n		
Flin	0.076	0.280		
Fnonlin	0.074	0.241		

[a] For C_{0h} = 1.700 mmol L^{-1}.

Before performing any comparative analysis of linear and non-linear forms of the three isotherm models, the best among the four linear Langmuir isotherm forms to be used in subsequent comparisons must be determined for adsorption of each Me species. For Cd(II) adsorption, Llin1 presents four EF_{min} (ARE, EABS, SAE and ADRSQ), Llin4 displays four too (RMSE, MPSD, ARSE and HYBRID), Llin3 exhibits one (CST) and Llin2 none, while SEPD value of Llin1 (13.65%) is smaller than that of Llin4 (24.91%), therefore, Llin1 is the selected form (Table 5). For Pb(II) adsorption, Llin4 is the best since it has six EF_{min} (ARE, EABS, SAE, RMSE, HYBRID and CST) compared with two (MPSD and ARSE) for Llin3, one (ADRSQ) for Llin1 and none for Llin2, as well as the lowest SEPD value (17.47%) (Table 6).

Subsequently, linear and non-linear forms of the same isotherm model are compared for adsorption of Cd(II) and Pb(II) (Figures 5 and 6, respectively). Regarding Langmuir isotherm, for Cd(II) adsorption, Llin1

is better than Lnonlin as the former has six EF_{min} (ARE, EABS, SAE, MPSD, ARSE and ADRSQ), whereas the latter has only three (RMSE, HYBRID and CST), and SEPD value of the former (20.42%) is smaller than that of the latter (36.54%); for Pb(II) adsorption, Llin4 is better than Lnonlin as indicated by seven EF_{min} (ARE, EABS, SAE, MPSD, ARSE, HYBRID and CST) versus two (RMSE and ADRSQ) and a smaller SEPD value (1153 versus 1377%). Concerning Freundlich isotherm, for Cd(II) adsorption, Fnonlin is better than Flin since it has six EF_{min} (ARE, EABS, SAE, RMSE, CST and ADRSQ) versus three (MPSD, ARSE and HYBRID) and a smaller SEPD value (1282 versus 1300%); for Pb(II) adsorption, Fnonlin compared with Flin presents six EF_{min} (ARE, EABS, SAE, RMSE, CST and ADRSQ) versus three (MPSD, ARSE and HYBRID) and a slightly larger SEPD value (1737 versus 1731%). As regards Temkin isotherm, for Cd(II) adsorption, Tlin and Tnonlin have equal error values and, consequently, identical EPD values and the same SEPD value (100.9%); for Pb(II) adsorption, Tlin and Tnonlin have practically the same error values and, as a consequence, very similar to each other (0 or very close to 0) EPD and SEPD values. It is noteworthy that, for adsorption of the two Me species, modeling results provided by linear regression are better (Langmuir isotherm) than or very similar (Temkin isotherm) to those offered by non-linear regression, which is in agreement with previously published data (Ho et al. 2002). Regression giving the best fit differs from one model to another, being linear, non-linear and both linear and non-linear for Langmuir, Freundlich and Temkin isotherms, respectively.

Then, a comparison is made among linear forms of all isotherms for adsorption of each Me species. For Cd(II) adsorption, Llin1 is the best with seven EF_{min} (ARE, EABS, SAE, MPSD, ARSE, HYBRID and ADRSQ), while Tlin has two (RMSE and CST) and Flin none, SEPD values for Llin1, Tlin and Flin being 20.42, 100.9 and 1300%, respectively (Figure 5). For Pb(II) adsorption, the best is Tlin, which has nine EF_{min} and the smallest SEPD value (0 versus 1153% for Llin4 and 1731% and Flin) (Figure 6).

Table 3. Values of error functions for Cd(II) adsorption

Error function[a]	Error function value								
	Langmuir isotherm					Temkin isotherm		Freundlich isotherm	
	Llin1	Llin2	Llin3	Llin4	Lnonlin	Tlin	Tnonlin	Flin	Fnonlin
ARE	**4.67**	5.36	5.69	4.90	5.26	5.62	5.62	11.19	11.04
EABS	**0.00282**	0.00346	0.00336	0.00289	0.00300	0.00285	0.00285	0.00673	0.00529
SAE	**0.0197**	0.0242	0.0235	0.0203	0.0210	0.0199	0.0199	0.0471	0.0371
RMSE	0.00371	0.00425	0.00369	0.00367	0.00363	**0.00350**	**0.00350**	0.00741	0.00612
MPSD	**7.77**	8.25	8.18	7.76	7.97	9.96	9.96	16.05	21.18
ARSE	**0.0710**	0.0753	0.0747	0.0708	0.0728	0.0910	0.0910	0.147	0.193
HYBRID	0.0332	0.0403	0.0329	0.0325	**0.0322**	0.0368	0.0368	0.124	0.138
CST	0.00187	0.00217	0.00170	0.00177	**0.00168**	0.00178	0.00178	0.00611	0.00529
ADRSQ	**0.997**	0.971	0.894	0.894	0.976	0.978	0.978	0.896	0.932

[a] For each error function, the bold value is E_0 for EPD calculation in Figure 5, i.e., the best among linear and non-linear form values of all isotherms (Llin1 is selected from the four linear Langmuir isotherm forms).

Table 4. Values of error functions for Pb(II) adsorption

Error function[a]	Error function value								
	Langmuir isotherm					Temkin isotherm		Freundlich isotherm	
	Llin1	Llin2	Llin3	Llin4	Lnonlin	Tlin	Tnonlin	Flin	Fnonlin
ARE	9.15	6.35	6.19	6.00	7.45	**3.45**	**3.45**	8.42	8.00
EABS	0.00327	0.00388	0.00353	0.00316	0.00332	**0.00171**	**0.00171**	0.00424	0.00326
SAE	0.0229	0.0272	0.0247	0.0221	0.0233	**0.0120**	**0.0120**	0.0297	0.0228
RMSE	0.00429	0.00474	0.00423	0.00403	0.00376	**0.00198**	**0.00198**	0.00462	0.00392
MPSD	16.21	8.98	8.75	9.13	11.06	**4.63**	4.64	10.96	14.60
ARSE	0.148	0.0820	0.0798	0.0834	0.101	**0.0423**	0.0424	0.100	0.133
HYBRID	0.0761	0.0494	0.0427	0.0423	0.0446	**0.00997**	0.00999	0.0541	0.0589
CST	0.00470	0.00248	0.00201	0.00190	0.00227	**0.00050**	**0.00050**	0.00269	0.00251
ADRSQ	0.998	0.984	0.912	0.912	0.950	**0.986**	**0.986**	0.942	0.946

[a] For each error function, the bold value is E_0 for EPD calculation in Figure 6, i.e., the best among linear and non-linear form values of all isotherms (Llin4 is selected from the four linear Langmuir isotherm forms).

Table 5. Error percent deviations (EPD) and EPD sums (SEPD) of linear Langmuir isotherm forms for Cd(II) adsorption

Isotherm model	EPD (%)									SEPD (%)
	ARE	EABS	SAE	RMSE	MPSD	ARSE	HYBRID	CST	ADRSQ	
Llin1	0	0	0	1.09	0.13	0.28	2.15	10.00	0	13.65
Llin2	14.78	22.70	22.84	15.80	6.31	6.36	24.00	27.65	2.61	143.1
Llin3	21.84	19.15	19.29	0.545	5.41	5.51	1.23	0	10.33	83.31
Llin4	4.93	2.48	3.05	0	0	0	0	4.12	10.33	24.91

Table 6. Error percent deviations (EPD) and EPD sums (SEPD) of linear Langmuir isotherm forms for Pb(II) adsorption

Isotherm model	EPD (%)									SEPD (%)
	ARE	EABS	SAE	RMSE	MPSD	ARSE	HYBRID	CST	ADRSQ	
Llin1	52.50	3.48	3.62	6.45	85.26	85.46	79.91	147.4	0	464.1
Llin2	5.83	22.78	23.08	17.62	2.63	2.76	16.78	30.53	1.40	123.4
Llin3	3.17	11.71	11.76	4.96	0	0	0.95	5.79	8.62	49.96
Llin4	0	0	0	0	4.34	4.51	0	0	8.62	17.47

Afterwards, non-linear forms of all isotherms for adsorption of the two Me species are compared. For Cd(II) adsorption, Lnonlin with five EF_{min} (ARE, MPSD, ARSE, HYBRID and CST) is better than Tnonlin with four (EABS, SAE, RMSE and ADRSQ) and Fnonlin with none; the same ranking of Lnonlin, Tnonlin and Fnonlin is indicated by SEPD values of 36.54, 100.9 and 1282%, respectively (Figure 5). For Pb(II) adsorption, all EF_{min} belong to Tnonlin, which also has the smallest SEPD value (0.66 versus 1377 and 1737% for Lnonlin and Fnonlin, respectively) (Figure 6).

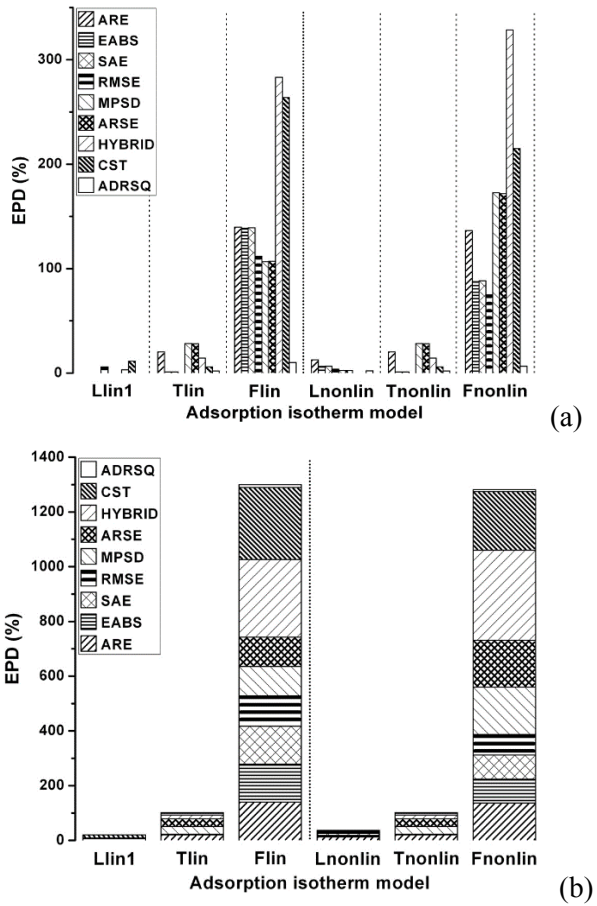

Figure 5. Error percent deviations (EPD) (a) and EPD sums (b) of linear and non-linear Langmuir, Temkin and Freundlich isotherms for Cd(II) adsorption.

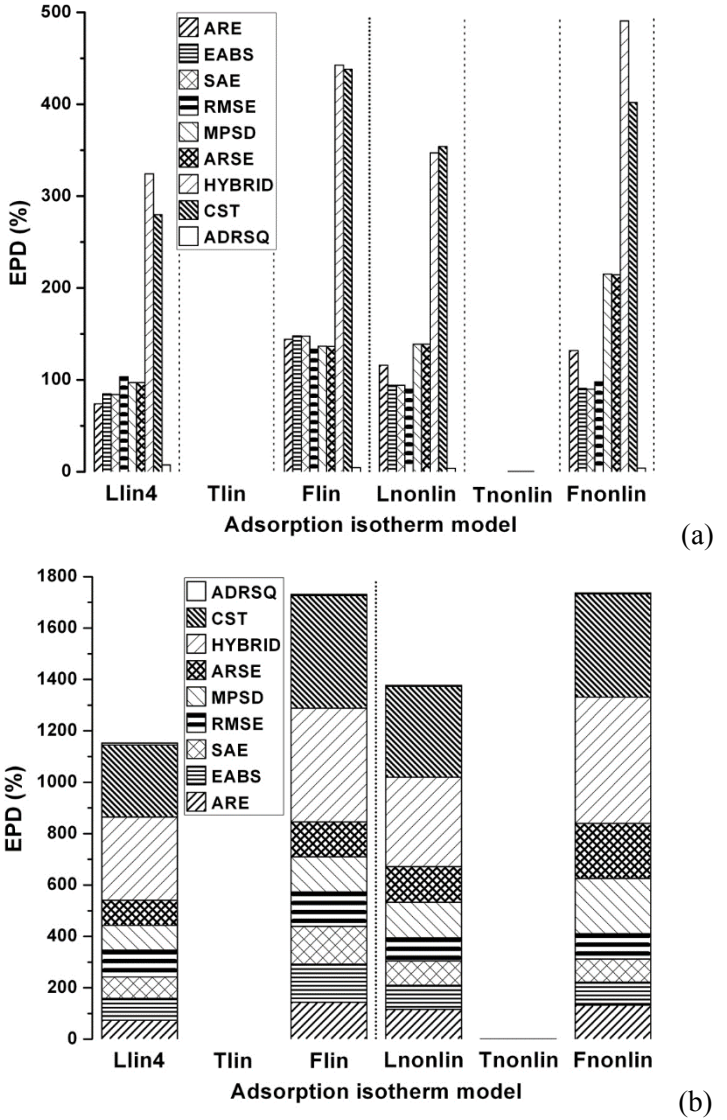

Figure 6. Error percent deviations (EPD) (a) and EPD sums (b) of linear and non-linear Langmuir, Temkin and Freundlich isotherms for Pb(II) adsorption.

The order of isotherm model validities for Cd(II) adsorption, indicated by both linear and non-linear regressions, is Langmuir > Temkin > Freundlich; the values of EPD (except those of ADRSQ) for Llin1,

Lnonlin, Tlin and Tnonlin are below 30%, whereas most of those for Flin and Fnonlin lie within the range of 135–330% (Figure 5). For Pb(II) adsorption, isotherm validity order revealed by using linear regression is the same with that indicated by non-linear regression, i.e., Temkin > Langmuir > Freundlich; EPD (except ADRSQ) values for Tlin and Tnonlin are equal or very close to 0, whereas those for Llin4, Lnonlin, Flin and Fnonlin range mostly from 130 to 490% (Figure 6). It is worth emphasizing that, for adsorption of both Me species, the descending order of isotherm model validities established by linear regression is identical with that determined via non-linear regression. It is noted that, among all linear and non-linear isotherm model forms considered, i.e., Llin1/Llin4, Lnonlin, Flin, Fnonlin, Tlin and Tnonlin, the highest validity is presented, for Cd(II) adsorption, by linear form Llin1 and, for Pb(II) adsorption, to practically the same extent by linear form Tlin and non-linear form Tnonlin.

The analysis of isotherm parameter values predicted by mathematical modeling gives useful information on adsorption of Cd(II) and Pb(II) (Tables 1 and 2, respectively). The values of q_m for Cd(II) and Pb(II) adsorption (0.112 and 0.074 mmol g^{-1}, respectively) are close to the highest corresponding q_e values (0.100 and 0.075 mmol g^{-1}, respectively); the q_m value for Cd(II) is larger than that for Pb(II), as expected considering q_e values. The binding energy towards Pb(II) is higher than that towards Cd(II), as indicated by the larger K_L value for the former Me species (20.52 L $mmol^{-1}$) compared with that for the latter (3.60 L $mmol^{-1}$). Temkin isotherm fits well (to similar extents when using linear and non-linear regressions) the experimental data, confirming that Me chemisorption takes place. Strong interactions Me–adsorbent consistent with chemisorption are indicated by the high values of Temkin parameters b_T and K_T (those estimated by linear regression are very close to the corresponding ones determined by non-linear regression for both Me adsorption); parameter values for Pb(II) are larger than the corresponding ones for Cd(II), pointing out that the forces binding the former Me species are stronger than those holding the latter, which is in agreement with what K_L values indicate (Zafar et al. 2007). Of the three models, Freundlich

isotherm gives the poorest fit to experimental data for each Me species, excluding the possibility that multilayer adsorption takes place and further confirming the occurrence of chemisorption that results in monolayer coverage of adsorbent surface (McKay 1995). The values of $1/n$ (0.312 and 0.241 for Cd(II) and Pb(II), respectively) are comprised within the range 0–1, showing favorable conditions for Me adsorption and therefore easy Me removal from aqueous solutions (Subramanyam and Das 2009; Hamdaoui and Naffrechoux 2007). The K_F value for Cd(II) is higher than that for Pb(II) (0.081 and 0.074 $mmol^{1-1/n}\ L^{1/n}\ g^{-1}$, respectively), which is in accordance with the larger q_m value of the former Me species compared with that of the latter. The values of ΔG^0 are negative for adsorption of both Me species (–19.96 and –24.20 kJ mol^{-1} for Cd(II) and Pb(II), respectively), indicating the feasibility and spontaneous nature of adsorption (Boparai et al. 2011). The R_L values (0.581 and 0.315 for Cd(II) and Pb(II), respectively) lying within the 0–1 range point out that adsorption is favorable, revealing that ILLF-SDVB beads constitute a good adsorbent for the two Me species.

CONCLUSION

Both linear and non-linear regressions were used in mathematical modeling of adsorption of the heavy metal ions Cd(II) and Pb(II) from aqueous solutions onto surface-functionalized polymer beads for comparatively analyzing the experimental equilibrium data by three isotherm models. The validities of Langmuir, Temkin and Freundlich models were evaluated by employing two criteria based on nine error functions. Langmuir and Temkin models successfully describe Cd(II) and Pb(II) adsorption, respectively; by contrast, Freundlich model gives the poorest description of adsorption of each Me species. It was evidenced that modeling results provided by linear regression may be better than or similar to those offered by non-linear regression. For adsorption of both Me species, the best fit to experimental data is obtained by using linear

regression, non-linear regression and both linear and non-linear regressions for Langmuir, Freundlich and Temkin models, respectively; the descending order of isotherm model validities determined by employing linear regression is the same with that established by using non-linear regression. Modeling results confirm that Me adsorption is a chemisorption process, revealing its feasibility and spontaneous nature, therefore point out that ILLF-SDVB beads have potential applications as adsorbent in wastewater treatment.

REFERENCES

Armagan, B.; Toprak, F. Optimum isotherm parameters for reactive azo dye onto pistachio nut shells: Comparison of linear and non-linear methods. *Polish Journal of Environmental Studies* 2013, 22, 1007–1011.

Bilba, N.; Bilba, D.; Moroi, G. Synthesis of a polyacrylamidoxime chelating fiber and its efficiency in the retention of palladium ions. *Journal of Applied Polymer Science 2004, 92, 3730–3735.*

Bilba, D.; Moroi, G.; Bilba, N. Copper (II) and mercury (II) retention properties of a polyacrylamidoxime chelating fiber. *Environmental Engineering and Management Journal* 2006, 5, 297–305.

Bilba, D.; Bilba, N.; Moroi, G. *Removal of mercury*(*II*) *ions from aqueous solutions by the* polyacrylamidoxime chelating fiber. *Separation Science and Technology* 2007, *42, 171*–184.

Boparai, H. K.; Joseph, M.; O'Carroll, D. M. Kinetics and thermodynamics of cadmium ion removal by adsorption onto nanozerovalent iron particles. *Journal of Hazardous Materials* 2011, 186, 458–465.

Bradl, H. Adsorption of heavy metal ions on clays. In *Encyclopedia of surface and colloid science update supplement*; Editor, P. Somasundaran; Marcel Dekker Inc.: New York, 2004; Vol. 5, pp. 35–47.

Brdar, M. M.; Takači, A. A.; Šćiban, M. B.; Rakić, D. Z. Isotherms for the adsorption of Cu(II) onto lignin – comparison of linear and non-linear methods. *Hemijska Industrija* 2012, 66, 497–503.

Casas, J. S.; Sordo, J. *Lead: chemistry, analytical aspects, environmental impact and health effects* (1st ed.); Elsevier: Amsterdam, 2006.

Foo, K. Y.; Hameed, B. H. Insights into the modeling of adsorption isotherm systems. *Chemical Engineering Journal* 2010, 156, 2–10.

Freundlich, H. M. F. Über die adsorption in lösungen. *Zeitschrift für Physikalische Chemie* 1906, 57A, 385–470. [Adsorption in solution. *Journal of Physical Chemistry* 57A: 385–470].

Hamdaoui, O.; Naffrechoux, E. Modeling of adsorption isotherms of phenol and chlorophenols onto granular activated carbon: Part I. Two-parameter models and equations allowing determination of thermodynamic parameters. *Journal of Hazardous Materials* 2007, 147, 381–394.

Han, R.; Zhang, J.; Han, P.; Wang, Y.; Zhao, Z.; Tang, M. Study of equilibrium, kinetic and thermodynamic parameters about methylene blue adsorption onto natural zeolite. *Chemical Engineering Journal* 2009, 145, 496–504.

He, J.; Hong, S.; Zhang, L.; Gan, F.; Ho, Y. S. Equilibrium and thermodynamic parameters of adsorption of Methylene Blue onto rectolite. *Fresenius Environmental Bulletin* 2010, 19, 2651–2656.

Ho, Y. S.; McKay, G. Sorption of dye from aqueous solution by peat. *Chemical Engineering Journal* 1998, 70, 115–124.

Ho, Y. S.; Porter, J. F.; McKay, G. Equilibrium isotherm studies for the sorption of divalent metal ions onto peat: copper, nickel and lead single component systems. *Water, Air, and Soil Pollution* 2002, 141, 1–33.

Kumar, K. V. Comparative analysis of linear and non-linear method of estimating the sorption isotherm parameters for malachite green onto activated carbon. *Journal of Hazardous Materials* 2006, B136, 197–202.

Kumar, K. V.; Porkodi, K.; Rocha, F. Isotherms and thermodynamics by linear and non-linear regression analysis for the sorption of methylene

blue onto activated carbon: Comparison of various error functions. *Journal of Hazardous Materials* 2008, 151, 794–804.

Langmuir, I. The constitution and fundamental properties of solids and liquids. Part I. Solids. *Journal of the American Chemical Society* 1916, 38, 2221–2295.

Legates, D. R.; McCabe, G. J. Jr. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydro-climatic model validation. *Water Resources Research* 1999, 35, 233–241.

McKay, G.; Blair, H. S.; Gardener, J. R. Adsorption of dyes on chitin. I. Equilibrium studies. *Journal of Applied Polymer Science* 1982, 27, 3043–3057.

McKay, G. (Ed.), *Use of Adsorbents for the Removal of Pollutants from Wastewaters;* CRC Press: Boca Raton, 1995.

Moroi, G.; Bilba, D.; Bilba, N. Thermal behaviour of palladium complexing polyacrylamidoxime polymer. *Polymer Degradation and Stability* 2001, 72, 525–535.

Moroi, G.; Bilba, D.; Bilba, N. Thermal degradation of mercury chelated polyacrylamidoxime. *Polymer Degradation and Stability* 2004, 84, 207–214.

Moroi, G.; Bilba, D.; Bilba, N.; Ciobanu, C. Thermal behaviour of polyacrylamidoxime-copper chelates. *Polymer Degradation and Stability* 2006, 91, 535–540.

Moroi, G. N. Investigation on structure and properties of cobalt(II)/polyesterurethane metallopolymer films. *Journal of Polymer Research* 2012, 19, 1–10.

Moroi, G. N.; Avram, E.; Bulgariu, L. Adsorption of heavy metal ions onto surface-functionalised polymer beads. I. Modelling of equilibrium isotherms by using non-linear and linear regression analysis. *Water, Air, and Soil Pollution* 2016, 227, 1–18. Erratum to: Adsorption of heavy metal ions onto surface-functionalised polymer beads. I. Modelling of equilibrium isotherms by using non-linear and linear regression analysis. *Water, Air, and Soil Pollution* 2016, 227, 1–2.

Salarirad, M. M.; Behnamfard, A. Modeling of equilibrium data for free cyanide adsorption onto activated carbon by linear and non-linear

regression methods. *2011 International Conference on Environment and Industrial Innovation IPCBEE,* 2011; 12, 79–84, IACSIT Press, Singapore.

Sigel, A.; Sigel, H.; Sigel, R. K. O. *Cadmium: from toxicity to essentiality*; Springer, Dordrecht, 2013.

Subramanyam, B.; Das, A. Linearized and non-linearized isotherm models comparative study on adsorption of aqueous phenol solution in soil. *International Journal of Environmental Science and Technology* 2009, 6, 633–640.

Temkin, M. I. Adsorption equilibrium and the kinetics of processes on nonhomogeneous surfaces and in the interaction between adsorbed molecules. *Zhurnal Fizicheskoi Khimii* 1941, 15, 296–332.

Wasewar, K. L. Adsorption of metals onto tea factory waste: A review. *International Journal of Research and Reviews in Applied Sciences* 2010, 3, 303–322.

Zafar, M. N.; Nadeem, R.; Hanif, M. A. Biosorption of nickel from protonated rice bran. *Journal of Hazardous Materials* 2007, 143, 478–485.

INDEX

A

absorption spectroscopy, 61
adjusted coefficient of determination, 160
adsorbent-adsorbate affinity, 151, 156
adsorption, v, vii, ix, 33, 34, 51, 53, 54, 59, 61, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178
adsorption isotherm modeling, ix, 149, 150
adsorption isotherms, 151, 176
adsorption mechanism, 151
algorithm, 15, 55, 93, 94
analytical applications, 98
aqueous solutions, 150, 154, 174, 175
Arrhenius equation, 35
average absolute error, 159
average relative error, 159
average relative standard error, 160

B

binding energy, 156, 157, 173
bioavailability, 15, 57

C

cadmium, 96, 154, 175
calibration, 3, 4, 7, 14, 15, 16, 17, 23, 34, 35, 36, 48, 49, 50, 52, 54, 55, 56, 57, 58, 59, 60, 61, 64, 65, 66, 67, 68, 86, 96, 97, 98, 99, 100, 101, 102, 103, 105, 107, 110, 111, 112, 113, 114
calibration programs, 23
Cd(II) adsorption, 162, 163, 165, 166, 167, 168, 170, 171, 172
chemical, viii, 3, 11, 55, 59, 67, 69, 102
chemical properties, 11
chemisorption, 157, 162, 173, 175
chemometrics, 71
chi-square test, 152, 160
chromatography, 15, 17, 35, 50, 57, 58, 67
coefficient of determination, 80, 86, 121, 125, 126, 133, 160
coefficient of variation, 14, 78, 110

D

data analysis, 46, 118
data generation, 129

data set, ix, 6, 7, 16, 37, 102, 106, 118, 121, 132
Dubinin–Radushkevich, 33, 52, 151
dynamic thermogravimetric analysis, 60

E

Environmental Protection Agency (EPA), 22, 54
enzyme-linked immunosorbent assay, 34, 67
enzyme(s), 15, 16, 18, 33, 34, 51, 54, 65, 67
equilibrium, vii, ix, 31, 32, 33, 34, 53, 61, 150, 151, 154, 156, 157, 158, 160, 162, 174, 176, 177, 178
equilibrium adsorption capacity, 151, 154, 156, 157, 158
equilibrium adsorption isotherms, 151
error functions, vii, ix, 150, 151, 152, 153, 155, 159, 161, 162, 168, 169, 174, 177
error percent deviations, 170, 171, 172

F

formula, 29, 74, 81, 127
Freundlich isotherm model, 155, 157, 161
function estimation, 11, 12, 16, 55, 65

G

Galileo, 96
Gibbs free energy change, 156

H

health effects, 176
heavy metal ions, vii, ix, 149, 150, 174, 175, 177
heterogeneity, 27, 158
heterogeneous variances, 27
heteroscedasticity, 38, 110
homogeneity, 3, 27, 28, 38, 39, 63, 114
hybrid fractional error function, 160

I

independent variable, vii, ix, 24, 87, 150, 160
intervals, 76
interval-valued, vii, ix
ionic liquid-like functionalities, 150
ions, 10, 175, 177
isotherm model parameters, 162, 166
isotherm model validity, ix, 149, 151
isotherm models, vii, ix, 33, 149, 150, 151, 152, 153, 154, 155, 161, 162, 166, 174, 178
isotherms, 151, 162, 163, 164, 165, 167, 168, 169, 171, 172, 177

K

kinetic constants, 53, 54
kinetic equations, 34, 53
kinetic parameters, 33
kinetics, 3, 14, 16, 34, 35, 57, 61, 65, 178
Koble–Corrigan, 151, 153

L

Lagrange multipliers, 87, 103, 111
Langmuir isotherm model, 155
LC-MS, 15, 55, 109
LC-MS/MS, 15, 55
least squares, viii, ix, 1, 2, 3, 4, 8, 14, 16, 17, 19, 25, 26, 46, 48, 52, 53, 54, 56, 60, 63, 64, 65, 68, 70, 73, 74, 77, 86, 88, 94, 96, 101, 103, 105, 107, 108, 109, 111, 114, 115, 118, 120, 121, 123, 124, 127, 128, 135, 142, 146, 147, 153

least squares method, weighting, transforming data, 2
least-squares regression analysis, 158
linear function, 78
linear model, 23, 25, 86, 126, 146
linear regression, v, vii, ix, 1, 2, 4, 5, 15, 16, 17, 20, 22, 23, 32, 33, 45, 47, 48, 49, 57, 58, 59, 61, 63, 64, 67, 70, 73, 82, 94, 95, 96, 101, 102, 106, 110, 112, 113, 114, 117, 118, 119, 120, 121, 123, 125, 126, 127, 128, 129, 131, 133, 135, 137, 139, 141, 143, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 167, 173, 174, 177
liquid chromatography, 3, 14, 15, 17, 51, 63, 65, 68

M

mass spectrometry, 14, 15, 17, 51, 57, 58, 67
mathematical modeling, v, ix, 149, 151, 155, 158, 173, 174
mercury, 101, 115, 175, 177
metal ion, vii, ix, 149, 150, 174, 175, 176, 177
metals, 16, 59, 60, 150, 178
methylene blue, 153, 176, 177
models, vii, viii, ix, 2, 3, 4, 11, 13, 14, 15, 22, 23, 33, 34, 36, 47, 54, 55, 58, 61, 66, 70, 80, 81, 85, 92, 95, 96, 99, 100, 101, 102, 105, 106, 107, 113, 114, 117, 118, 121, 128, 129, 130, 131, 132, 135, 146, 148, 149, 150, 151, 152, 153, 154, 155, 158, 161, 162, 166, 173, 174, 176, 178
Monte Carlo method, 16, 65

N

National Bureau of Standards, 60
neglect, 105
Netherlands, 56
nickel, 176, 178
nicotinamide, 20
NOAA, 132
non-linear regression, vii, ix, 26, 149, 150, 151, 152, 153, 154, 167, 172, 173, 174, 176, 178
normal distribution, 27, 37, 95

O

orthogonal regression, vii, viii, 70, 90, 92, 112

P

palladium, 175, 177
Pb(II) adsorption, 155, 158, 161, 162, 164, 165, 166, 167, 169, 170, 171, 172, 173, 174
polymeric supports, 150
polymer(s), vii, ix, 33, 53, 149, 150, 174, 177

Q

quality control, 15, 64
quantification, 15, 34, 61, 68

R

radius, 119, 120, 122, 123, 131, 134, 135
random errors, 8, 9, 62, 101, 104
Redlich– Peterson isotherm model, 152
regression, vii, viii, ix, 1, 2, 3, 4, 5, 8, 15, 16, 17, 20, 22, 23, 26, 29, 34, 35, 45, 46, 48, 49, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 63, 64, 65, 68, 70, 71, 73, 76, 77, 80, 81, 82, 83, 84, 85, 86, 87, 89, 90, 92, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111,

112, 113, 114, 115, 117, 118, 120, 121, 123, 124, 126, 128, 129, 135, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 158, 160, 167, 173, 174, 176, 177, 178
regression analysis, ix, 2, 3, 8, 15, 56, 59, 64, 101, 102, 111, 145, 149, 151, 152, 158, 176, 177
regression equation, vii, ix, 48, 107, 108, 150, 160
regression line, vii, viii, 64, 70, 76, 77, 97, 99, 108, 129
regression method, x, 94, 96, 110, 150, 153, 178
regression model, vii, viii, 15, 29, 34, 35, 56, 61, 63, 64, 70, 86, 95, 97, 104, 105, 106, 109, 121, 126, 146, 147
root mean square error, 159
root(s), 23, 24, 28, 47, 79, 97, 159

S

sample variance, 124, 131
science, 47, 98, 105, 175
scientific theory, 47
separation factor, 156
set theory, 118
simple linear regression, 20, 33, 120
solution, 14, 36, 63, 82, 106, 124, 127, 128, 142, 151, 154, 156, 157, 158, 176, 178
sorption, 16, 33, 35, 61, 176
sorption process, 33
spectrophotometry, 3, 56, 62
statistics, 4, 5, 10, 60, 71, 80, 85, 86, 102, 105, 113, 118
styrene-divinylbenzene copolymer beads, 150
sum of squared errors, 155
sum of the absolute errors, 159
surface properties, 151
surface-functionalized polymer beads, vii, ix, 149, 150, 174

T

Temkin isotherm model, 157
thermodynamic parameters, 176
thermodynamics, 175, 176
transformation(s), viii, 1, 2, 3, 7, 9, 22, 23, 24, 25, 27, 28, 29, 30, 31, 33, 35, 36, 37, 40, 42, 43, 44, 46, 47, 48, 49, 50, 51, 53, 55, 56, 58, 59, 60, 61, 62, 63, 67, 87, 120
treatment, 34, 47, 48, 89, 150, 175
trigonometric functions, 23

U

universal gas constant, 157

V

validation, 15, 47, 50, 58, 61, 64, 99, 102, 105, 113, 114, 115, 131, 133, 177
vapor, 42, 43, 44, 45, 46
variables, viii, 1, 2, 3, 6, 7, 11, 23, 67, 70, 78, 83, 92, 94, 95, 98, 100, 102, 103, 111, 115, 145, 147, 148
variations, 15, 32
volatile organic compounds, 57

W

weighted regression, 15, 16, 56, 60, 92, 100, 103